Taxing Humans: Pitfalls of the Mechanism Design Approach and Potential Resolutions

Alex Rees-Jones and Dmitry Taubinsky*

October 23, 2017

Abstract: A growing body of evidence suggests that psychological biases can lead different implementations of otherwise equivalent tax incentives to result in meaningfully different behaviors. We argue that in the presence of such failures of "implementation invariance," decoupling the question of optimal feasible allocations from the tax system used to induce them—the "mechanism design approach" to tax analysis—cannot be the right approach to analyzing optimal tax systems. After reviewing the diverse psychologies that lead to failures of implementation invariance, we illustrate our argument by formally deriving three basic lessons that arise in the presence of these biases. First, the mechanism design approach neither estimates nor bounds the welfare computed under psychologically realistic assumptions about individuals' responses to the tax instruments used in practice. Second, the optimal allocations from abstract mechanisms may not be implementable with concrete tax policies, and vice-versa. Third, the integration of these biases may mitigate the importance of information asymmetries, resulting in optimal tax formulas more closely approximated by classical Ramsey results. We conclude by proposing that a "behavioral" extension of the "sufficient statistics" approach is a more fruitful way forward in the presence of such

*Rees-Jones: The Wharton School, University of Pennsylvania and NBER (E-mail: alre@wharton.upenn.edu). Taubinsky: University of California, Berkeley and NBER (E-mail: dtaubinsky@gmail.com). We are grateful to Robert Moffitt and Daniel Reck for helpful comments and advice.

psychological biases.

1 Introduction

A standard assumption in optimal tax policy design is that individuals' behavior is governed only by the choice-sets induced by the tax system—conditional on the choice-set induced, behavior does not vary across the tax systems that could be used to implement that choice set. This assumption—which we refer to as *implementation invariance*—reduces the question of optimal tax-system design to an optimization problem over a set of feasible consumption bundles satisfying incentive compatibility and government revenue constraints. The abstraction from the practical considerations of tax policy implementation results in a framework that is tractable and fruitful. This "mechanism design approach" to taxation has been broadly applied to characterize the features of optimal policy in both static (e.g., Mirrlees, 1971) and dynamic settings (for a review, see Golosov *et al.*, 2007).

In this paper, we articulate a challenge to the practical value of this approach: due to the psychological complexity of how individuals respond to taxation, the details of the tax system that induces a given choice set can substantially influence the resulting behavior. The growing evidence on the prevalence of taxpayer confusion, of heuristic optimization, and of imperfect attention suggests that the assumption of implementation invariance does not hold in practice. When this assumption fails, a policy analyst can be lead awry by the common two-step procedure of first considering the incentives induced by the optimal mechanism and only later considering its implementation.

In section 2, we summarize a series of recent empirical demonstrations of confusion, inattention, and heuristic use, all of which lead people to suboptimally respond to tax incentives. For each class of biases, we illustrate concretely the violations of implementation invariance that result. We argue that biases in the understanding of taxes are widespread, that these biases affect central economic behaviors, and that these biases are shaped by the idiosyncrasies

of different tax mechanisms in complex and subtle ways.

In section 3, we formalize the consequences of violations of implementation invariance for normative tax analysis. We build on a simple two-type model of optimal income taxation proposed by Stiglitz (1982) in which individuals choose between pairs of before-tax income z (which corresponds to a choice of labor supply) and the resulting after-tax consumption c. Using several behavioral biases as examples, we formalize three implications. First, the presence of these biases prevents the application of the revelation principle, a core tool of mechanism design that allows the analyst to separate the question of analyzing optimal behavior under a "direct mechanism" from the specifics of the tax which implements it. As a result, welfare under a direct mechanism neither estimates nor bounds the welfare attainable at the true optimal policy. Second, we illustrate that there are biases that can render the optimal allocation in a direct mechanism unimplementable with taxes, while the allocation resulting from the optimal taxes is unimplementable with a direct mechanism. Third, we show that the presence of these biases can mitigate the role of information rents—a central concept of mechanism design—and can ultimately result in tax analysis that more closely resemble that of frameworks that are not tightly centered on understanding information asymmetries, such as the Ramsey approach.

In section 4, we assess the comparative advantages of alternative approaches to tax policy analysis in the presence of psychological biases. We argue that a modification to the "sufficient statistics" approach (for a review, see Chetty, 2009) provides a fruitful way forward. This approach works with an allowable set of tax instruments directly, deriving optimal tax formulas involving elasticities and empirically-estimable formulations of bias. In addition to its tractability, this approach also transparently highlights deviations from standard optimal tax formulas. We present and discuss the key challenges to this approach, and discuss its comparative advantages to the mechanism design approach.

2 Violations of implementation invariance

To focus ideas and define basic concepts, we begin by discussing a recent experiment that cleanly demonstrates a failure of implementation invariance. We then turn to a series of examples demonstrating this phenomenon, and its causes, in the field.

2.1 A stylized lab example

The cleanest possible demonstration of a violation of implementation invariance would consist of a comparison of behavior under two meaningfully different tax instruments that induce the same choice sets. Tax policies in the field are rarely deployed in a manner that offers this comparison directly. However, as pursued in Abeler & Jäger (2015), labor markets may be designed in a laboratory setting that exactly satisfy these constraints.

Abeler & Jäger create a simple approximation to a labor-supply decision within the microcosm of the lab. The participants in their experiment must decide how much labor to provide in order to fund consumption. Labor is measured in the context of a real-effort task adopted from Gill & Prowse (2012), in which the participant may move a series of hundred-point slider scales to prespecified values. When time expires, participants receive a piece-rate wage for each slider that is positioned on its assigned value. This experimental task is arguably tedious, but it provides the participant with a means to trade current leisure for experimental earnings.

In the experiment, earnings from this task are subject to a progressive tax. Across treatment arms, the experimenters apply two tax systems that induce similar choice sets, but are of significantly differing complexity. In the "simple treatment", the progressive tax is implemented with two simply-articulated rules. The tax schedule traced by these rules can be calculated with relatively little effort. In the "complex treatment", the progressive tax is implemented with 22 tax rules. The tax schedule traced by these rules closely approximates that in the "simple treatment"—and thus induces approximately the same choices sets—but

the calculation of this tax schedule is substantially more cognitively demanding.

While the traditional mechanism design approach would treat these experimental taxes as interchangeable tools for achieving the same behavior, Abeler & Jäger document substantially different behavior across treatment arms. When nearly identical tax incentives were induced through the complex system, subjects were less likely to choose the payoff-maximizing output level, and on average earned 23% less than subjects in the simple treatment arm. Furthermore, as new tax rules were introduced across rounds in the experiment, subjects were systematically less responsive to tax changes in the complex frame. In short, these mechanisms had differing effects on the distortionary impact of taxation, despite the near equivalence of the choice-sets that the policies induced.

The mechanism design approach takes as given that we may use arbitrarily complex tools to induce the choice sets, and thus choices, that the mechanism designer views as desirable. In practice, however, the quality of decision-making might decline if the choice environment is imperfectly understood. This worry is undoubtedly relevant for behavior in the current U.S. income tax system, commonly lamented for its extraordinary complexity.

2.2 Field evidence

Laboratory experiments such as those of Abeler & Jäger provide compelling illustrations of failures of implementation invariance, but do not inform us about the biases that shape people's responses to the actual tax systems used in practice. We now discuss the evidence on biased responses to tax incentives in the field. We focus on biases caused by confusion, by heuristic adoption, and by differential salience of different tax provisions.

2.2.1 Confusion

Perhaps the most straightforward and psychologically uncomplicated manner in which psychological realism might influence our tax policy analysis is through the serious treatment of confusion. If a taxpayer misunderstands the provisions of the tax, he will come to be-

lieve he faces a different choice-set than he actually does. Under such circumstances, even an otherwise-optimizing agent would appear to generate violations of implementation invariance if the details of a tax instrument affect the manner in which it might be misunderstood.

Given the dramatic complexity of taxes in the United States, it is perhaps unsurprising that substantial confusion regarding tax provisions has been documented. When directly surveyed about the key parameters characterizing their federal income tax burden—like their marginal tax rate—taxpayers regularly report values with substantial individual error (Fujii & Hawley, 1988; Blaufus *et al.*, 2013; Gideon, 2015; Rees-Jones & Taubinsky, 2016). Analysis of observational data reveals that there are large differences in knowledge of taxpayers' understanding of the tax code: Chetty *et al.* (2013) find significant differences in bunching at the refund-maximizing kink of the earned income tax credit (EITC), and show that individuals who move from low-bunching neighborhoods to high bunching neighborhoods increase their EITC refunds due to new information diffusion. Moreover, significant amounts of tax benefits are "left on the table" every tax year through, e.g., failures to claim itemized deductions (Benzarti, 2016) or failures to claim the EITC (Bhargava & Manoli, 2015). The difficulty individuals face in understanding the complex tax code is argued to have generated the large professional-tax-preparation industry in the United States (Slemrod & Bakija, 2008). Attempts to "teach the tax code" have been shown to be ineffective, on average, but can be effective when paired with expert advice (as in, e.g., Chetty & Saez, 2013).

To concretely illustrate the potential for confusion to generate violations of implementation invariance, we focus on recent evidence arising from the work of Feldman *et al.* (2016). The authors present a clear test of the possibility that taxpayers mistake predictable changes in lump-sum transfers for changes in marginal tax incentives.

Feldman *et al.* examine the effect of the Child Tax Credit (CTC), a transfer given to households with a child younger than 17 in the calendar year. While the size of this transfer varies with income, virtually all filers with adjusted gross income between $30,000 and $100,000 were able to claim the maximum $1,000 credit in the window studied by the

authors. For this group, a loss of the CTC constitutes a lump-sum change in tax liability.

The requirement that a household have a child under 17 at the end of the calendar year introduces a discontinuity in the average tax credit received. A household whose child "ages out" on December 31, 2010 could not claim the CTC for 2010, whereas a household whose child "ages out" on January 1, 2011 could. This distinction is perfectly predictable. Furthermore, the distinction does not change the marginal tax rate, and thus should not influence marginal tax incentives except through small income effects. However, using a regression discontinuity design, the authors document that the loss of the CTC is associated with an approximately 0.5 percent decline in reported wage income relative to households who have just retained the credit for another year. The authors document that this effect is not driven by strategic timing of earnings, nor by direct effects of a child aging. They interpret their result as evidence that at least some households confuse factors that influence average tax rates with those that determine marginal tax rates.

Under the assumption that households with a child born in late December do not have meaningfully different preferences than those with a child born in early January, these results illustrate a clear violation of implementation invariance. Given that the CTC does not mechanically affect marginal tax rates for the group studied, the loss of this credit does not meaningfully induce different tradeoffs between lesiure and consumption. But by nevertheless changing labor supply, the CTC must therefore have shaped taxpayers' misunderstanding of the tax system. Upon observing an increase in their tax bill, the taxpayer incorrectly infers that marginal tax rates have gone up, and changes choice behavior.

2.2.2 Heuristic Adoption

As documented by a large literature in psychology, decision makers often adopt simple heuristics to approximate complex decision-rules when cognitively effortful decisions must be quickly and regularly made. In an influential paper, Liebman & Zeckhauser (2004) consider and formalize two heuristics that they argue are sensible, and potentially common, means

of approximating a convex schedule like the US income tax. These heuristics are presented in figure 1, and are described below. Our notation and summary of these heuristics draws heavily from our prior work on their empirical measurement (Rees-Jones & Taubinsky, 2016), which we also summarize below.

The first heuristic, *ironing*, is applied by individuals who know the average tax rate they face, and forecast tax liability by applying their average tax rate to all incomes. Using the ironing heuristic, the forecasted tax at income z is given by $\tilde{T}_I(z|z^*, \theta) = A(z^*|\theta) * z$, where z^* denotes the individual's own income, θ denotes all individual-specific characteristics that determine the applicable tax schedule, and $A(z^*|\theta)$ denotes the individual's average tax rate. This heuristic has the practical benefit that it leads to reasonably accurate beliefs about the *levels* of taxes when considering small deviations from one's current income. Thus, for decisions about how to budget one's annual income, this heuristic leads to minimal errors.

However, when used to infer the leisure/consumption combinations that form an individual's choice set, this heuristic leads to meaningful errors. Specifically, it leads to overestimation of the tax burden for comparatively low incomes and underestimation of the tax burden for comparatively high incomes. This heuristic directly generates inaccurate beliefs about marginal tax rates: because the tax schedule is convex, average tax rates are systematically smaller than marginal tax rates, and thus the application of this heuristic generates a "flattening" of perceived schedules.

Feldman *et al.* (2016) argue that this heuristic potentially generates the confusion over marginal tax rates they document, and similar responsiveness to shocks to average tax rates have been documented in lab settings (de Bartolome, 1995). In a recent survey experiment directly eliciting perceptions of tax schedules, Rees-Jones & Taubinsky (2016) find evidence that the ironing heuristic is adopted by 29-43% of US tax filers. In a non-tax application, Ito (2014) shows that this heuristic rationalizes consumer response to non-linear electricity pricing schedules.

The second heuristic, *spotlighting*, is applied by individuals who know their own tax and

own *marginal* tax rate, and forecast tax liability by applying their marginal rate to the difference between their own income and the income amount under consideration. Using the spotlighting heuristic, the forecasted tax at income z is given by $\tilde{T}_S(z|z^*,\theta) = T(z^*|\theta) + MTR(z^*|\theta) * (z - z^*)$, where z^* again denotes the individual's own income, $MTR(z^*|\theta)$ denotes the marginal tax rate at that income, and $T(z^*|\theta)$ denotes the true tax due at that income.

Within one's own tax bracket, the spotlighting heuristic leads to correct beliefs about the level and slope of the tax schedule; as a result, under the assumption that leisure/consumption pairs falling under other tax brackets are irrelevant alternatives in the choice set, this heuristic does not meaningfully violate implementation invariance. While this heuristic has received some theoretical attention, Rees-Jones & Taubinsky (2016) find little evidence of its adoption in their forecasting experiment.

2.2.3 Salience

A recent and growing literature has demonstrated that the visibility of taxes substantially influences behavioral response, and that this feature can be incorporated into standard tax formulas with appropriate care. In two pioneering studies, Chetty *et al.* (2009) demonstrated that experimentally manipulated integration of taxes into posted prices for groceries and alcohol meaningfully influenced the resulting demand curves, and Finkelstein (2009) demonstrated that the reduced visibility of road-use tax induced by the adoption of "EZ-pass" reduced taxes' (dis)incentive effect on road use. Other recent advancements have studied how issues of salience affect the regressivity of commodity taxes (Goldin & Homonoff, 2013), how a social planner would optimally choose between differentially salient tax instruments (Goldin, 2015), and how issues of endogenous salience might affect tax policy analysis (Feldman *et al.*, 2015; Taubinsky & Rees-Jones, Forthcoming). In short, in the context of commodity taxation, salience has come to be viewed as an increasingly well-understood instrument that shapes the welfare evaluation of tax policy.

While the tax salience literature has often focused on commodity taxation, its core findings appear to apply to tax incentives administered through the income tax as well. Miller & Mumford (2015) examine a change to the Child and Dependent Care Credit (CDCC) introduced in 2003; this change affected the direct, visible value that could be claimed for this credit that, considered in isolation, increased the subsidization of child and dependent care administered through the income tax. However, this policy change also interacted with provisions of the existing Child Tax Credit in a non-salient but offsetting manner, in many cases resulting in a net decrease in subsidization when all interactions are taken into account. Miller & Mumford demonstrate that taxpayer response was most consistent with reaction to the salient direct incentives of the tax, and with complete ignorance of the arguably non-salient interactions with other provisions of the tax code. As summarized by the authors, "taxpayers increased their expenditure on child care in response to the expansion of the CDCC regardless of whether the actual after-tax price of child care increased or decreased."

Under reasonably mild assumptions on the demand for child care, these results imply a violation of implementation invariance. For consumers facing an increase in the after-subsidy price of childcare, the larger amount of childcare demanded post-reform was available in their choice set prior to the tax change. Under the assumption that this change in subsidies does not introduce implausibly large income effects, this necessitates a violation of our key assumption. As a concrete illustration of the failure of implementation invariance, one may consider the predicted effect of a completely transparent price change as contrasted with the price-change introduced through interactions with multiple price interactions. If one believes that transparently raising the price of child care would lower its demand—i.e., that child care is not a Giffen good—then behavior under these two choice-set-equivalent instruments would vary.

2.3 Summary

While we have emphasized several classes of psychological response that generate failures of implementation invariance, we note that this list is not exhaustive. Even in the absence of psychological biases, similar issues can arise when otherwise-similar tax systems induce different compliance costs on taxpayers. In the domain of psychological responses, we have not reviewed the substantial literature studying the influence of heuristics and biases on tax compliance decisions,[1] instead focusing our attention on demonstrations directly tied to the labor supply and commodity demand issues that are central to welfare analysis.

Even in the settings of greatest policy interest, our empirical understanding of taxpayer psychology remains highly incomplete. Absent a complete theory that specifies, e.g., the formation of misperceptions and the adoption or rejection of biases, it is not possibly to fully characterize the situations in which the failure of implementation invariance will become first-order. These caveats aside, across the examples we have considered, we find empirical support for the notion that implementation invariance fails in several field settings of direct policy interest. Furthermore, we present evidence that misperceptions of the tax schedule are sufficiently widespread that such failures could conceivably extend to any decisions that rely on accurate calculation of marginal tax incentives. While further work is needed to trace the limits of when these failures do, and do not arise, the current literature suggests these issues arise often enough that attention to their theoretical consequences is warranted.

3 Consequences of the failure of implementation invariance

In this section, we illustrate several key consequences of the failure of implementation invariance for the formal analysis of tax policy. While the lessons we present are quite general, we

[1] For recent field evidence, see Engström *et al.* (2015) and Rees-Jones (2017). For a broader review, see Kirchler & Braithwaite (2007).

illustrate these lessons in the context of a standard, but simple, two-type model of income taxation. We proceed by presenting the model, describing the mechanism-design approach to its analysis, and then illustrate the key complications that arise when biases depend on tax instruments.

3.1 A standard optimal income tax model

We consider a simple model of income taxation based on Stiglitz (1982). There are two "types" of individuals in the economy, indexed by their earnings ability$\theta \in \{L, H\}$. Those of low earnings ability ($\theta = L$) earn a wage $w(L)$ per unit of labor and those of high earnings ability ($\theta = H$) earn a wage $w(H)$ per unit of labor, where $w(H) > w(L)$. The fraction of each type in the population is denoted by $\alpha(\theta)$. An agent with wage w generates gross income of $z = w \cdot l$ when he supplies l units of labor. All post-tax income is spent on consumption c which, together with the labor output l, generates utility $U(c, l)$. This utility is typically assumed to be concave, increasing in consumption, and decreasing in labor. In some of the analysis that follows, we make the simplifying assumption that $U(c, l) = c - \psi(l)$, where $\psi', \psi'' > 0$.

The government's objective is to maximize social welfare:

$$W = \sum_{\theta} \alpha(\theta) \cdot G(U(c(\theta), l(\theta))) \tag{1}$$

We assume that the government's evaluation of individual utility, G, is a smooth and concave function.

The policy decision faced by the government is to specify a tax-and-transfer system that maximizes social welfare. The assumption that G is concave reflects the government's disfavor of inequality, and thus the optimal tax system would redistribute income from those with high earnings ability to those with low earnings ability. Ability is not observed, however, and so the tax must depend on the signal of ability contained in observable earnings ($z(\theta)$).

In contrast to taxing ability, taxing earnings is distortionary. When those with high earnings are taxed and those with low earnings are subsidized, a high-earnings-ability worker might choose to reduce his labor supply in order to represent himself as a low-earnings-ability worker.

3.2 A two-step approach to solving the optimal tax problem

This formulation of the social welfare problem illustrates a key trade-off in tax policy design. On the one hand, the tax system must redistribute income from those of high earnings ability to those of low earnings ability. On the other hand, this tax system must simultaneously account for the fact that such redistribution can lead workers to misrepresent their ability type through the earnings that they choose. Simultaneously mathematically accommodating both the policy motives of the government and the misrepresentation motives of the individual can be challenging. However, a powerful result from mechanism design—the revelation principle—can dramatically simplify the necessary analysis, and forms the heart of what we term "the mechanism design approach" to tax policy.

The revelation principle, as originally articulated in Myerson (1979), states that any equilibrium allocation that can arise among fully optimizing agents can be achieved as an equilibrium allocation in a *direct mechanism*—that is, a mechanism that induces agents to truthfully report their type. This allows analysis to be divided into two simplified steps: first, characterizing behavior in a world where agents are incentivized to report their type, and second, characterizing the tax system that induces those incentives. We illustrate these two steps in the context of our simple model below.

Step 1: characterizing the direct mechanism. Rather than assuming that the government only observes earnings, now assume that agents "announce" their type, $\theta \in \{L, H\}$. The planner assigns an allocation that depends on that announcement, $(z(\theta), c(\theta))$. The set of allocations must satisfy incentive compatibility (IC) constraints—which ensure that individuals are incentivized to announce their types honestly—and a budget balance (B)

constraint—which ensures that total consumption in the economy does not exceed total earnings. Formally, the government maximizes

$$\max_{(c(\theta),z(\theta))} \sum_{\theta} \alpha(\theta) \cdot G(U(c(\theta), \frac{z(\theta)}{w(\theta)})) \tag{2}$$

subject to the constraints

$$c(H) - \psi(z(H)/w(H)) \geq c(L) - \psi(z(L)/w(H)) \quad \text{(IC-}H\text{: no incentive for } H \text{ types to lie)}$$
$$c(L) - \psi(z(L)/w(L)) \geq c(H) - \psi(z(H)/w(L)) \quad \text{(IC-}L\text{: no incentive for } L \text{ types to lie)}$$
$$z(H) + z(L) \geq c(H) + c(L) \quad \text{(B)}$$

Typically, only conditions IC-H and B are binding at the optimum. If high-ability taxpayers are indifferent between their allocation and the allocation of the low-ability taxpayers, then low-ability taxpayers will strictly prefer the allocation that entails less consumption since generating income is more costly for them.

Step 2: implementing the direct mechanism. Once the optimal direct mechanism is characterized, the second step is to reverse-engineer the tax system that would implement the incentives in that optimum. In the simple optimal taxation model presented here, this is straightforward. The income tax function must satisfy $T(z(\theta)) = z(\theta) - c(\theta)$, and it must assign sufficiently high punishments to deviations from earning $z(H)$ or $z(L)$. A smooth tax function would, for example, have to satisfy $(1 - T'(z(\theta)))U_c(c(\theta), z(\theta)/w(\theta)) + \frac{1}{w(\theta)}U_l(c(\theta), z(\theta)/w(\theta)) = 0$ to ensure that individuals do not want to deviate from their assigned allocations $(c(\theta), z(\theta))$. Generally, while the optimal direct mechanism is unique, it can be implemented with many different kinds of tax functions.

3.3 Implementation invariance and its failure

In the context of this simple model, we may define implementation invariance as a restriction that taxpayers' preferences over consumption and labor cannot be influenced by the step-two tax system induced. Consider an individual who chooses a consumption-earnings bundle (c, z) over (c', z') when both options are available. This decision is implementation invariant if any tax system T satisfying $z - T(z) = c$ and $z' - T(z') = c'$ results in the same apparent preference. The literature reviewed in the previous section suggests violations of this principle arise in situations where inattention, misperception, or heuristics guide decisions.

Notice that individuals whose decisions are not implementation invariant violate basic tenants of optimization appealed to in the statement of the revelation principle. As a result, use of the two-stage procedure in the previous section is no longer ensured to be valid. This failure may be understood to be generated by a disjoint between the incentive-compatability constraints that restrict a fully-optimal decision maker and the *perceived* incentive-compatibility constraints that govern a biased decision-maker. Stated informally, the incentive-compatability constraint generates a threshold on "how much" you can tax an individual before inducing a misrepresentation of type. If different tax systems generate different types of misunderstanding, then they similarly generate different such thresholds. This complicates analysis, but also introduces new tools to the policy maker.

It is worth noting that many commonly-studied biases do not operate through this channel of misunderstood incentive compatibility constraints. For example, behavioral models of prospect theory or sophisticated present bias are better understood as cases where the decision maker does accurately understand the constraints faced, but holds an individual utility function that is viewed as normatively undesirable by the social planner (e.g., attending to "irrelevant" reference comparisons or applying impatient time discounting). Cases such as these need not generate violations of implementation invariance; indeed, variants of the mechanism design approach have been successfully applied to these biases (see, e.g., Kanbur *et al.*, 2008; Lockwood, 2015).

3.4 Consequences of the failure of implementation invariance

We use a series of examples to illustrate several key implications of the violation of implementation invariance. In each example, we assume that choices reveal preferences under a direct mechanism, whereas choices may be biased when incentives must be inferred from a tax. While the contrasts and consequences that we highlight are exposited under highly stylized assumptions, we believe they illustrate the the broader point that the welfare analysis of tax policies can lead to meaningfully different conclusions in the presence of this class of biases. In cases where the results are not proven in-text, proofs may be found in the Appendix.

Lesson 1: The optimal tax system may induce a consumption-labor allocation that is different than the one implemented with the optimal direct mechanism. The allocation induced by the optimal tax system may generate higher or lower welfare than would be induced under the direct mechanism.

To demonstrate Lesson 1, consider the consequences of the salience of an income tax. Suppose that when individuals choose labor supply, they make decisions based on a perception of the tax represented by $\tilde{T} = \sigma T$. When $\sigma = 1$, individuals correctly attend to the taxes in place. When $\sigma > 1$, taxes are overly salient. When $0 \leq \sigma < 1$, taxes are partially ignored.

To illustrate the impact of salience on welfare, consider first the extreme case where individuals choose labor supply as if there is no tax in place ($\sigma = 0$). In this case, the tax is entirely ignored, and as a result it does not distort behavior: regardless of the tax, individuals choose the efficient level of labor supply satisfying $\psi'(l(\theta)) = w(\theta)$. This means that it is possible to achieve full redistribution without creating inefficiencies, simply by choosing a tax function that satisfies $z(H) - T(z(H)) = z(L) - T(z(L))$. In constrast, under the approach taken in the mechanism design problem of section 3.2, this first-best level of labor supply would be viewed as unobtainable. The presence of this bias facilitates the maximization of our social welfare function.

In contrast, when taxes are overly salient ($\sigma > 1$), the distortionary consequences of a

tax are even greater than they would be under the assumption of optimal behavior. Since distortionary motives are the primary cost of redistribution, in this case the presence of this bias hinders the maximization of our social welfare function.

This may be summarized in the following formal result:

Proposition 1. *At the optimal tax system, the social welfare function expressed in equation 1 is decreasing in scaling parameter σ. When $\sigma < 1$, the welfare that results in the optimal tax system is higher than would be obtained under the optimal direct mechanism. When $\sigma > 1$, the welfare that results in the optimal tax system is lower than would be obtained under the optimal direct mechanism.*

Lesson 2: The allocation implemented by the optimal direct mechanism may not be implementable by any income tax. Conversely, equilibrium allocations obtainable under some biases may not be implementable by a direct mechanism among optimizers.

We illustrate this point by a simple example of a taxpayer who adopts the ironing heuristic. As reviewed in section 2.2.2, this taxpayer perceives the tax schedule to be linear, with slope $\tau(z(\theta)) = T(z(\theta))/z(\theta)$. Further suppose that $\psi(l) = l^2/2$.

Under the direct mechanism, the binding IC constraint is given by

$$\text{c(H)-c(L)}=\frac{z(H)^2 - z(L)^2}{2w(H)^2} \quad \text{(Direct Mechanism IC)} \tag{3}$$

Under ironing, the misperception of the tax schedule leads to the different first-order condition $1 - T(z(\theta))/z(\theta) = z(\theta)/w(\theta)^2$. Since $c(\theta) = z(\theta) - T(z(\theta))$, this implies that $c(\theta)/z(\theta) = z(\theta)/w(\theta)^2$, and thus that $c(\theta) = z(\theta)^2/w(\theta)^2$. Thus under ironing, the consumption allocations must satisfy

$$\text{c(H)-c(L)}=\frac{z(H)^2}{w(H)^2}-\frac{z(L)^2}{w(L)^2} \quad \text{(Ironing IC)} \tag{4}$$

Generically, it cannot be the case that $\frac{z(H)^2}{w(H)^2}-\frac{z(L)^2}{w(L)^2} = \frac{z(H)^2-z(L)^2}{2w(H)^2}$, which implies that the optimal bundle under the direct revelation mechanism cannot be implemented with an income tax when individuals iron.

Conversely, because in the two-type model ironing leads the high wage types to think that the tax on low wage types is higher than it actually is, the low type's actual labor-consumption allocation would appear much more appealing to the high types were it shown in a direct mechanism. This suggests that the direct mechanism frame may provide a higher incentive for the high types to deviate downward. Proposition 2 below provides a more general characterization.

Proposition 2. *Consider the social welfare function in equation 1 and suppose individuals are ironers. Generically, there does not exist a tax function T that implements the optimal allocation of the optimal direct mechanism. Moreover, the resulting allocation of consumption that is obtained from solving for the optimal tax function T cannot be implemented using a direct mechanism when $\psi(l) = l^\rho/\rho$ for $\rho > 1$ sufficiently small.*

This leads to the broader lesson that the set of allocations that are are feasible when taxpayers are perfect optimizers might not be feasible when considering taxpayers' imperfect reactions to "real-world" policy tools. Conversely, desirable "real-world" outcomes may seem infeasible when analyzed under the assumption of perfect optimization.

Lesson 3: The reaction to information asymmetries that generates the key tension of the mechanism design approach may be mitigated or eliminated.

Recall that in the standard model, perfect redistribution is not possible because the high type must have incentives that are high enough to not imitate the low type. With $\psi(l) = l^2/2$, this

incentive compatibility constraint is presented in equation (3). The constraint captures the key innovation of optimal tax analysis in the spirit of Mirrlees (1971): because of asymmetric information, taxes can still be distortionary even without any "arbitrary" constraints on the tax tools such as linearity. The optimal taxation problem thus builds on broader principles of mechanism design of maximizing transfers from the high types by paying them minimal "information rents."

Misperceptions of taxes can fundamentally change the principles of optimal tax analysis, and may completely eliminate the role of concepts such as "information rents." Indeed, this outcome has already been demonstrated when discussing Lesson 1 above, in which distortionary behavior was eliminated in the case where perceived taxes were scaled to zero. Intuitively, these findings mirror the growing set of demonstrations that behavioral biases can mitigate the negative consequences of information asymmetries in insurance markets, for the similar reason that agents cannot claim rent for information that they have ignored (Handel, 2013; Handel & Kolstad, 2015; Handel *et al.*, 2015; Spinnewijn, 2017). While the assumptions of the illustration in Lesson 1 are extreme, more generally the impact of heuristics and biases can be to mitigate the role of information rents and to push optimal tax analysis more towards the mechanics represented in models of Ramsey taxation.

We illustrate this idea by demonstrating the reversal of a core principle of taxation: that in the presence of income taxation, commodity taxes should only be used if they help to target taxes to those of high earnings ability.

Consider, following Stiglitz (1982), an extension of the model insection 3.1, in which individuals choose before-tax income z and a consumption bundle (c_1, c_2). One interpretation is that c_1 and c_2 are different commodities. Another interpretation is that c_1 is period 1 consumption and c_2 is period 2 consumption. For simplicity, assume that $U(c_1, c_2, l, \theta) = u(c_1) + v(c_2, \theta) - \psi(l)$.

In the standard model, when both types L and H have the same subutility $v(c_2, \theta) \equiv v(c_2)$, the optimal allocation must always satisfy $v'(c_2(\theta)) = u'(c_1(\theta))$ for each type (Stiglitz,

1982). This means that linear, nonlinear, or means-tested taxes on c_2 are *not* justified when different types' preferences are homogeneous. This result is not specific to a two-type model and holds more generally for a continuum of types (Atkinson & Stiglitz, 1976; Saez, 2002; Golosov *et al.*, 2013).

For unbiased consumers, taxes (or subsidies) on c_2 are justified only when they can be used to better screen between low and high types. When those of higher earnings ability have a greater preference for c_2 (i.e., $\frac{v_{c_2}(c_2,\theta)}{u'(c_1)}$ is increasing in θ), it then becomes optimal to have some form of a tax on c_2.[2] Greater consumption of c_2 now serves as an additional signal that an individual has high earnings ability, and thus taxing these individuals can efficiently increase the redistributive properties of the tax system. Explicit formulas for optimal taxes on c_2 are complex, however, as they depend intricately on the informational advantages that the commodity taxes have over the income tax.

The case for commodity taxation can be fundamentally affected by the presence of more realistic psychological assumptions. In particular, the psychological assumption that individuals perfectly compute the labor-supply incentives induced by commodity taxes is quite demanding; more realistically, consumers might at least partially neglect the labor-supply incentives induced by taxes on c_2.

To illustrate formally, suppose the government chooses an income tax $T(z)$ on before-tax earnings and a linear commodity tax t on c_2. The individual first chooses earnings z and a consumption bundle c_1 and c_2 such that $c_1 + (1+t)c_2 \leq z - T(z)$. Suppose, however, that individuals neglect to consider the tax t on c_2 when choosing their labor supply, and only react to the commodity tax after they have generated their income and are observing the after-tax prices of both c_1 and c_2. Letting $g(\theta)$ denote the social marginal utility of income to a type θ, the effects of increasing the commodity tax are now as follows:

- A decrease in revenue following a substitution away from c_2, given by $t\frac{dc_2}{dt} = -t\zeta\frac{\bar{c}_2}{1+t}dt$, where $\bar{c}_2$ denote aggregate consumption of c_2 and ζ is the price elasticity of (aggregate)

[2] Conversely, when higher types have a lower preference for c_2, it is optimal to have some form of a subsidy on c_2 (Atkinson & Stiglitz, 1976; Saez, 2002; Golosov *et al.*, 2013).

demand for c_2.

- A mechanical revenue effect given by $\bar{c}_2 dt$, where $\bar{c}_2$ denotes total consumption of c_2.
- A mechanical welfare effect given by $-E[g(\theta)c_2(\theta)]dt$.

The sum of these effects must be zero at the optimum:

$$-\lambda t\zeta \frac{\bar{c}_2}{1+t}dt + \lambda \bar{c}_2 dt - E[g(\theta)c_2(\theta)]dt = 0.$$

where λ is the marginal value of public funds. Solving the above equation for t then yields the following result:

Proposition 3. *When individuals are inattentive to the commodity tax on the labor supply margin, the optimal commodity tax t satisfies*

$$\frac{t}{1+t} = \frac{\lambda - E[g(\theta)\tilde{c}_2(\theta)]}{\lambda\zeta} \tag{5}$$

where $\tilde{c}_2(\theta) = c_2(\theta)/\bar{c}_2$ is the share of c_2 consumption by type θ, and λ is the marginal value of public funds.

There are several noteworthy features of formula (5). First, notice that it is the standard Ramsey formula with redistributive concerns (Diamond, 1975). Second, notice that the formula holds regardless of the extent to which preferences for c_2 differ between high and low types: whether the Engel curve for c_2 is driven by income effects or heterogeneous preferences correlated with earnings ability does not matter. In contrast to the core lessons from mechanism design, the formula for the optimal commodity tax here does not depend at all on the extent to which introducing distortions to $\frac{v_{c_2}(c_2,\theta)}{u'(c_1)}$ allows the designer to reduce the information rents that must be payed to the high types. This is because individuals ignore the tax t on the labor supply margin, and thus the presence of the income tax does not fundamentally change the basic logic fleshed out in the classical Ramsey approach.

3.5 Broader implications for mechanism design

In this paper, we have been critical of what we've termed the "mechanism design approach", specifically referring to the 2-step procedure outlined in section 3.2. We do not mean to suggest that the general approach of mechanism design itself should be abandoned. Rather, when pursuing the goals of mechanism or policy design, we urge caution when making use of simplifying short-cuts that rely critically on perfect individual understanding and rationality in situations where behavioral biases are widespread, We have documented the consequences of these simplifying assumptions in the tax setting, but we note that similar tensions have been documented when analyzing consumers' failures to reveal their "valuation type"' in Becker-DeGroot-Marschak mechanisms (Cason & Plott, 2014) or auctions (Kagel *et al.*, 1987), or student's failure to reveal their "preference type" in matching mechanisms (Rees-Jones, Forthcoming; Hassidim *et al.*, 2016).

4 Discussion

If the manner in which taxes are implemented is fundamentally intertwined with the manner in which decisions are made, the two-stage procedure of separating the question of optimal behavior under direct mechanisms from the question of implementing the direct mechanism poses a difficult foundation for the integration of psychological realism. Instead, the computation of optimal feasible allocations and the implementation of these allocations must be considered simultaneously.

The simultaneous consideration of these two questions is implicit in the alternative approach summarized by Diamond & Saez (2011), which is to first write down a limited set of possible tax instruments and then to optimize over those instruments. Within this framework, a particularly fruitful technique has been to express optimal tax formulas in terms of measurable "sufficient statistics" such as elasticities or social marginal welfare weights. Because of the emphasis on measurable responses to actual tax instruments, this approach is

more easily extended to incorporate psychological biases. The key additional statistic needed to compute optimal tax policy is a price-metric measure of bias: a monetized measure of the difference between what people would optimally do and what they actually do.

4.1 A concrete illustration of the sufficient statistics approach

To provide a more concrete illustration of the sufficient statistics approach, we summarize the formula provided by Farhi & Gabaix (2015) for a nonlinear income tax with a continuum of productivity types, and for utility functions of the form $U(c,l) = c - \psi(l)$. In particular, assume that individuals perceive the actual income tax T to be $\tilde{T}$, where $\tilde{T}(z)$ depends on the actual income tax $T(z)$ on earnings z, as well as the the individual's actual earnings z^*, and the tax paid on those earnings $T(z^*)$.

This formulation captures both the salience and ironing examples studied in the previous section. In the case of salience, $\tilde{T}(z) = \sigma T(z)$. In the case of ironing, $\tilde{T}(z) = T(z^*) + (z - z^*)\frac{T(z^*)}{z^*}$.

Farhi & Gabaix (2015) show that for this broad class of misperceptions, the optimal tax rates depend on the sum of two terms. The first term, $\frac{T_R'(z)}{1-T_R'(z)}$, is the standard optimal tax formula for rational consumers, as characterized by Saez (2001). This depends on the governments' redistributive preferences as well as the usual measurable statistics: the distribution of earned income and the elasticity of taxable income with respect to the marginal tax rate.[3]

The second term, denoted $\tilde{\tau}^b(z)$, is essentially a price metric for consumers biases. This term answers the following question: if consumers were fully debiased, by what percent would the marginal keep rate, $1 - T'(z)$, need to be increased so that consumers choose the

[3]To define this term formally, let H be the cumulative density function of income, with a probability density h. Let ζ be the elasticity of taxable income with respect to the keep rate $1 - T'(z)$. Let h^* be the "virtual density" $h^*(z) := \frac{h(z)}{1-T'(z)+\zeta z T''(z)}$. Then the optimal income tax satisfies

$$\frac{T_R'(z)}{1-T_R'(z)} = \frac{1}{\zeta(z)} \frac{1-H(z)}{\lambda z h(z)} E[g(z')|z' \geq z]$$

where λ is the marginal value of public funds and $g(z)$ is the social marginal utility of income to a z-earner.

same amount of labor as they do in the biased state? Formally, $\tilde{\tau}^b(z) = \frac{(1-T'(z))-\frac{1}{2}\psi'(z/w)}{1-T'}$.

With these two terms in hand, Farhi & Gabaix (2015) show that the optimal income tax satisfies

$$\frac{T'(z)}{1-T'(z)} = \tilde{\tau}^b + \frac{T'_R(z)}{1-T'_R(z)} \tag{6}$$

Formula (6) provides an immediate characterization of the optimal income taxes for the salience and ironing biases we have discussed. In the case of salience, we have $\tilde{\tau}^b(z) = (1-\sigma)\frac{T'(z)}{1-T'(z)}$, which leads to the simple formula $\frac{T'(z)}{1-T'(z)} = \frac{1}{\sigma} \cdot \frac{T'_R(z)}{1-T'_R(z)}$. In the case of ironing, we have $\tilde{\tau}^b(z) = 1 - \frac{A(z)}{1-T'(z)}$, which can also be plugged into (6) to obtain a formula for the optimal income tax.

An under-appreciated insight is that while $\tilde{\tau}^b(z)$ could be the result of many different psychologies, the empirical strategy used to quantify $\tilde{\tau}$ does not have to depend on the psychology in play, and can be largely an extension of standard revealed preference methods. Once the "welfare-relevant domain" (Bernheim & Rangel, 2009) is identified, the bias measure is constructed as the wedge between choices in the welfare relevant domain and the choices normally observed. This approach has been applied to assess the welfare costs of biases in a variety of tax settings, such as quantifying the consequences of salience (see, e.g., Chetty *et al.*, 2009; Taubinsky & Rees-Jones, Forthcoming). In these approaches, the authors compute the change in posted prices that would alter demand as much as a debiasing intervention that displays tax-inclusive final prices—the assumed welfare relevant domain.

A simple example of empirically quantifying such price-metrics in a non-income tax domain is provided by Allcott & Taubinsky (2015), who run an experiment that provides a direct estimate of bias for each consumer's valuation of energy efficient lightbulbs (CFLs). They compute willingness to pay (WTP) for more versus less energy efficient lightbulbs in a standard market frame, and then measure how the distribution of WTP changes when biases arising from inattention or incorrect beliefs are eliminated via an informational intervention that directs attention.

While experiments of this nature are compelling, they place significant demands on empirical implementation: they rely on within-subject manipulation of the welfare frame[4] and observation of the consequences of that change on behavior holding all else constant. In some cases, experiments satisfying these desiderata are infeasible, such as when assessing labor supply response to income tax misperception. However, other strategies for measuring bias are still available. For example, Gerritsen (2016) measures the disjoint between chosen and "happiness"-maximizing labor supply based on subjective well-being data, and integrates the resulting wedge into optimal tax analysis. Rees-Jones & Taubinsky (2016) directly measure individual heuristic use in the context of a forecasting experiment, and then assess the predicted consequences of these biases if they were acted upon in a standard model of labor-supply determination.

Through the adoption of empirical strategies like those discussed above, sufficient statistics formulas such as those presented by Farhi & Gabaix (2015) are fully implementable using standard methods for estimating elasticities, and extensions of standard revealed preference methods for computing price-metric measures of bias.

4.2 Challenges for future work

An important challenge with extending the sufficient statistics approach to incorporate individuals' mistakes is the critical need to have individual-level measures of biases, rather than just population means. In the context of Allcott and Taubinsky's (2015) welfare analysis of taxes on inefficient lightbulbs, the effects of a tax change are determined not by the population average of bias, but rather by the bias of consumers who are marginal to a tax change. Allcott and Taubinsky show that without restrictive assumptions, the only way to obtain the necessary measures of bias is to estimate bias at the individual level. In the context of Taubinsky and Rees-Jones's (Forthcoming) analysis of sales taxes in the presence of heterogeneous inattention, the mean and the variance of consumers' scaling of the tax, together

[4]E.g., manipulating the presentation of tax-inclusive and tax-exclusive prices, or examining CFL purchasing behavior before and after information provision.

with the standard components of rational deadweight-loss calculations, are necessary and sufficient for computing efficiency costs. Thus, in contrast to the case when bias is homogenous (Chetty *et al.*, 2009), aggregate data is insufficient for policy analysis: measurement of heterogeneity across individuals is necessary for understanding welfare effects.

These examples illustrate a key challenge that arises in economic welfare analysis when biased decision-makers are present. In a standard model of optimizing consumers, marginal benefits must equal the marginal costs or the price at the margin. Thus, even if consumers are heterogeneous, the marginal consumers who determine market prices have homongenous valuations. This "marginal homogeneity" is essential for inferring welfare by observing only aggregate changes in behavior. In the presence of behavioral biases, however, marginal benefits need not equal marginal costs, and this difference will be heterogeneous when consumers are heterogeneous in their biases. While analysts often restrict their attention to biases' impact on average incentives, the role of biases as hetereogeneity-inducing devices can be, at times, of greater quantiative importance to policy analysis.

While the presence of heterogeneity on the margin does complicate analysis, it does not fundamentally change the principles by which the standard sufficient statistics approach may be deployed, nor does it fundamentally change the strategies for how bias should be measured. However, applying these strategies in the presence of this heterogeneity does require especially rich data sets that allow for robust measurement at the individual level. Observational or quasi-experimental data of this type is not always available, requiring researchers to design new experiments that allow more granular measurement. As the literature progresses, an iterative application of the sufficient statistics approach to welfare, paired with granular measurement of heterogeneous biases in tax settings, appears to be both a conceptually justified and practically implementable approach to the development of empirically informed tax policy.

References

Abeler, Johannes, & Jäger, Simon. 2015. Complex Tax Incentives. *American Economic Journal: Economic Policy*, **7**(3), 1–28.

Allcott, Hunt, & Taubinsky, Dmitry. 2015. Evaluating Behaviorally-Motivated Policy: Experimental Evidence from the Lightbulb Market. *American Economic Review*, **105**(8), 2501–2538.

Atkinson, Anthony B, & Stiglitz, J.E. 1976. Design of Tax Structure - Direct Versus Indirect Taxation. *Journal of Public Economics*, **6**, 55–75.

Benzarti, Youssef. 2016. How Taxing Is Tax Filing? Leaving Money on the Table Because of Hassle Costs. *Working Paper*.

Bernheim, B Douglas, & Rangel, Antonio. 2009. Beyond Revealed Preference: Choice-theoretic Foundations for Behavioral Welfare Economics. *The Quarterly Journal of Economics*, **124**(1), 51–104.

Bhargava, Saurabh, & Manoli, Day. 2015. Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment. *American Economic Review*, **105**(11), 3489–3529.

Blaufus, Kay, Bob, Jonathan, Hundsdoerfer, Jochen, Sielaff, Christian, Kiesewetter, Dirk, & Weimann, Joachim. 2013. Perception of Income Tax Rates: Evidence from Germany. *European Journal of Law and Economics*, **40**(3), 457–478.

Cason, Timothy N., & Plott, Charles R. 2014. Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing. *Journal of Political Economy*, **122**(6), 1235–1270.

Chetty, Raj. 2009. Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods. *Annual Review of Economics*, **1**, 451–488.

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

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

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

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

Diamond, Peter, & Saez, Emmanuel. 2011. The Case for a Progressive Tax: From Basic Research to Policy Recommendations. *The Journal of Economic Perspectives*, **25**(4), 165–190.

Diamond, Peter A. 1975. A Many-person Ramsey Tax Rule. *Journal of Public Economics*, **4**, 335–342.

Engström, Per, Nordblom, Katarina, Ohlsson, Henry, & Persson, Annika. 2015. Tax Compliance and Loss Aversion. *American Economic Journal: Economic Policy*, **7**(4), 132–64.

Farhi, Emmanuel, & Gabaix, Xavier. 2015. Optimal Taxation with Behavioral Agents. *NBER working paper No. 21524*.

Feldman, Naomi, Goldin, Jacob, & Homonoff, Tatiana. 2015. Raising the Stakes: Experimental Evidence on the Endogeneity of Taxpayer Mistakes. *Working paper*.

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

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

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

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

Gideon, Michael. 2015. Do Individuals Perceive Income Tax Rates Correctly? *Public Finance Review*.

Gill, David, & Prowse, Victoria. 2012. A Structural Analysis of Disappointment Aversion in a Real Effort Competition. *American Economic Review*, **102**(1), 469–503.

Goldin, Jacob. 2015. Optimal Tax Salience. *Journal of Public Economics*, **131**, 115 – 123.

Goldin, Jacob, & Homonoff, Tatiana. 2013. Smoke Gets in your Eyes: Cigarette Tax Salience and Regressivity. *American Economic Journal: Economic Policy*, **5**(1), 302–336.

Golosov, Mikhail, Tsyvinski, Aleh, & Werning, Ivan. 2007. New Dynamic Public Finance: A User's Guide. *Pages 317–388 of: NBER Macroeconomics Annual 2006, Volume 21*. NBER Chapters. National Bureau of Economic Research, Inc.

Golosov, Mikhail, Troshkin, Maxim, Tsyvinski, Aleh, & Weinzierl, Matthew. 2013. Preference Heterogeneity and Optimal Capital Income Taxation. *Journal of Public Economics*, **97**, 160–175.

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

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

Handel, Benjamin R., Kolstad, Jonathan T., & Spinnewijn, Johannes. 2015. Information Frictions and Adverse Selection: Policy Interventions in Health Insurance Markets. *NBER Working paper No. 21759.*

Hassidim, Avinatan, Romm, Assaf, & Shorrer, Ran. 2016. 'Strategic' Behavior in a Strategy-Proof Environment. *SSRN Working Paper No. 2784659.*

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

Kagel, John H., Harstad, Ronald M., & Levin, Dan. 1987. Information Impact and Allocation Rules in Auctions with Affiliated Private Values: A Laboratory Study. *Econometrica*, **55**(6), 1275–1304.

Kanbur, Ravi, Pirttilä, Jukka, & Tuomala, Matti. 2008. Moral Hazard, Income Taxation and Prospect Theory. *Scandinavian Journal of Economics*, **110**(2), 321–337.

Kirchler, E., & Braithwaite, V. 2007. *The Economic Psychology of Tax Behaviour.* Cambridge University Press.

Liebman, Jeffrey B., & Zeckhauser, Richard. 2004. Schmeduling. *Working Paper.*

Lockwood, Benjamin. 2015. Optimal Income Taxation with Present Bias. *Working Paper.*

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

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

Myerson, Roger B. 1979. Incentive Compatibility and the Bargaining Problem. *Econometrica*, **47**(1), 61–73.

Rees-Jones, Alex. 2017. Quantifying Loss-Averse Tax Manipulation. *The Review of Economic Studies*, rdx038.

Rees-Jones, Alex. Forthcoming. Suboptimal Behavior in Strategy-Proof Mechanisms: Evidence from the Residency Match. *Games and Economic Behavior*.

Rees-Jones, Alex, & Taubinsky, Dmitry. 2016. Heuristic Perceptions of the Income Tax: Evidence and Implications for Debiasing. *Working Paper*.

Saez, Emmanuel. 2001. Using elasticities to derive optimal income tax rates. *The Review of Economic Studies*, **68**(1), 205–229.

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

Slemrod, Joel, & Bakija, Jon. 2008. *Taxing Ourselves, 4th Edition: A Citizen's Guide to the Debate over Taxes*. The MIT Press.

Spinnewijn, Johannes. 2017. Heterogeneity, Demand for Insurance, and Adverse Selection. *American Economic Journal: Economic Policy*, **9**(1), 308–43.

Stiglitz, Joseph E. 1982. Self-Selection and Pareto Efficient Taxation. *Journal of Public Economics*, **17**(2), 213–240.

Taubinsky, Dmitry, & Rees-Jones, Alex. Forthcoming. Attention Variation and Welfare: Theory and Evidence from a Tax Salience Experiment. *Review of Economic Studies*.

A Appendix Proofs

A.1 Proof of Proposition 1

We show that welfare under the optimal income tax is decreasing in σ. This will establish the whole proposition since for $\sigma = 1$, consumers perceive the taxes correctly and thus the

direct mechanism formulation is equivalent to the the optimal income tax formulation.

For a given σ, let T_σ be the optimal income tax. Note that it must bind at the high types' IC constraint:

$$z(H) - \sigma T_\sigma(z(H)) - (z(L) - \sigma T_\sigma(z(L)) = \psi(z(H)/w(H)) - \psi(z(L)/w(H)).$$

We now show that if σ decreases by some small amount $d\sigma$, it will be possible to achieve more redistribution while keeping the before-tax choices $z(H)$ and $z(L)$ constant. To that end, define $T_{\sigma-d\sigma}$ to satisfy $(\sigma - d\sigma)T_{\sigma-d\sigma}(z) = \sigma T_\sigma(z)$; that is, $T_{\sigma-d\sigma}(z) = \frac{\sigma}{\sigma-d\sigma}T_\sigma(z)$. Then, because $T_\sigma(z(H)) > 0$ and $T_\sigma(z(L)) < 0$ at the optimum, this tax must achieve more redistribution, which increases social welfare. And by construction, it still satisfies the high types' IC constraint:

$$z(H)-(\sigma-d\sigma)T_{\sigma-d\sigma}(z(H))-(z(L)-(\sigma-d\sigma)T_{\sigma-d\sigma}(z(L)) = \psi(z(H)/w(H))-\psi(z(L)/w(H)).$$

This new allocation must increase welfare since the labor earnings are held constant while the distribution of consumption becomes more equal.

A.2 Proof of Proposition 2

First we show that the consumption bundle from the optimal direct mechanism can't be implemented with an income tax amongst ironers. The optimal direct mechanism must satisfy the classic "no distortion at the top" result: $\psi'(l(H)) = w_H$. Now the only way an income tax can implement the same result for the H types is by putting no tax on those consumers; if they do see a positive tax $T(z(H))$, then their average tax rate will be positive, which will make them think that their marginal tax rate is positive because of the ironing heuristic, and so they will want to choose labor satisfying $\psi'(l(H)) < w(H)$. But since the tax raises no money from the high types, it must then also satisfy $T(z(L)) = 0$. Thus, if

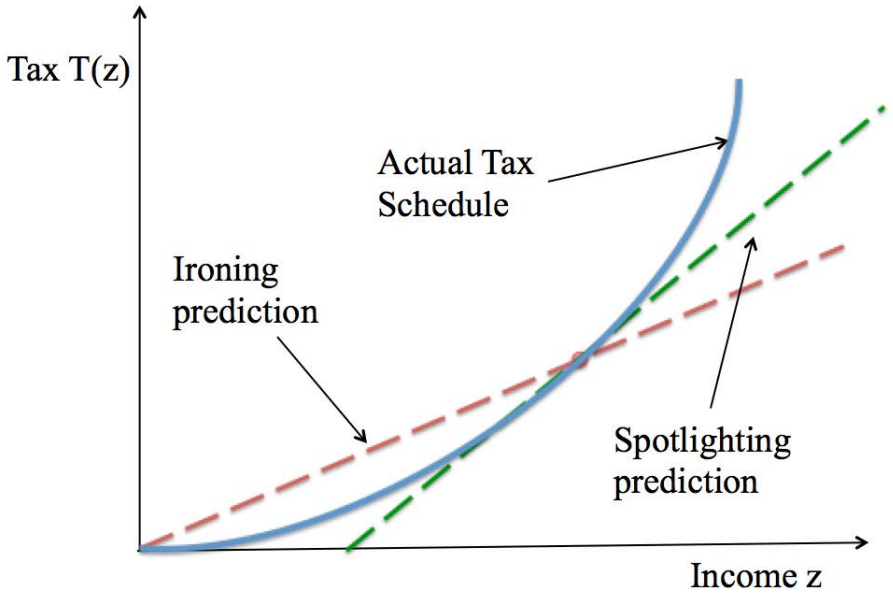
the income tax does not distort the high types' labor earnings, then it cannot achieve any redistribution. However, the optimal direct mechanism achieves a partially redistributive allocation satisfying $c(H) < z(H)$ and $c(L) > z(L)$. To summarize, the optimal direct mechanism achieves some redistribution, while generating no distortion to the high types' labor earnings. With the ironing psychology, however, it is impossible to satisfy both of these criteria.

Second, we show that the consumption bundle from the optimal tax cannot be implemented with an optimal direct mechanism. When individuals are ironers, the first order condition is $\frac{\psi'(z(\theta)/w(\theta))}{w(\theta)} = 1 - T(z(\theta))/z(\theta)$. But since $c(\theta) = z(\theta) - T(\theta)$, this implies that $c(\theta) = \psi'(z(\theta)/w(\theta)) \cdot (z(\theta)/w(\theta))$. Thus for this allocation to be implementable in a direct mechanism under the assumption that $\psi(l) = l^\rho/\rho$, it must satisfy

$$(z(H)/w(H))^\rho/\rho - (z(L)/w(H))^\rho/\rho \leq (z(H)/w(H))^\rho - (z(L)/w(L))^\rho.$$

For ρ sufficiently close to 1, however, the left-hand side of the above inequality will actually be greater than the right-hand side because $z(L)/w(H) < z(L)/w(L)$.

Figure 1: Heuristics for approximating a tax schedule



Notes: This figure presents an illustration of the ironing and spotlighting heuristics applied to a generic convex schedule. When using these heuristics, the taxpayer linearizes the convex schedule according to parameters local to his own position on the schedule, indicated by the red dot. Under the ironing heuristic, the taxpayer forecasts by applying his average tax rate at all points, resulting in the observed secant line. 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 line.
Source: Rees-Jones & Taubinsky (2016).