When Do "Nudges" Increase Welfare?

Hunt Allcott, Daniel Cohen, William Morrison, and Dmitry Taubinsky*

September 20, 2023

Abstract

We use public finance sufficient statistic approaches to characterize the welfare effects of "nudges," such as simplified information and warning labels, in markets with endogenous prices and government-set taxes. While many studies focus on average effects, we show that welfare also depends on how the nudge affects the variance of choice distortions, and average effects become irrelevant with zero pass-through or optimal taxes. We implement the framework with new experiments evaluating automotive fuel economy labels and sugary drink health labels. In both experiments, labels increase ostensibly beneficial behavior, but may decrease welfare because they increase the variance of choice distortions.

JEL Codes: D12, D61, D83, D90, H21, I18, L15, Q58.

Keywords: Behavioral welfare analysis, nudges, information provision, warning labels, randomized experiments, fuel economy, sugar-sweetened beverages.

*Allcott: Stanford and NBER. allcott@stanford.edu. Cohen: Northwestern University. daniel.cohen@kellogg.northwestern.edu. Morrison: Berkeley. wmorrison@berkeley.edu. Taubinsky: Berkeley and NBER. dmitry.taubinsky@berkeley.edu. We thank Victoria Pu for exceptional research assistance. We thank Anna Grummon, Christina Roberto, Cass Sunstein, and seminar participants at the Geneva School of Economics and Management, the NBER Roybal Conference, the NBER Summer Institute, New York University, Pompeu Fabra, Stanford GSB, Stanford Institute for Theoretical Economics, UC Berkeley, and the University of Copenhagen for helpful comments. We thank the National Science Foundation, the Alfred P. Sloan Foundation, National Institute on Aging (via the NBER Roybal Center grant #P30AG034532), and Time Sharing Experiments for the Social Sciences for grant funding. The experiment was registered in the American Economic Association Registry for randomized trials (available from www.socialscienceregistry.org/trials/7460) and at ClinicalTrials.gov (https://clinicaltrials.gov/ct2/show/NCT05038163), and was approved by Institutional Review Boards at NYU (protocol number FY2020-3805) and Berkeley (#2020-08-13558). Replication files are available from allcott.stanford.edu/research/.

In recent years, policymakers and academics have become increasingly interested in "nudges": policy instruments that can change behavior without directly changing prices or choice sets, such as simplified information provision, warning labels, reminders, social comparisons, framing, and choice architecture. Spurred by the book Nudge (Thaler and Sunstein 2008) and the success of the Behavioral Insights Team in the UK, more than 200 public agencies now incorporate nudges into the policy process (OECD 2017). Hundreds of academic papers evaluate nudges using randomized experiments (Delmas, Fischlein and Asensio 2013; Hummel and Maedche 2019), and the U.S. government alone has conducted more than 160 randomized evaluations of nudges (DellaVigna and Linos 2022). Nudges are being used to increase retirement savings, healthy eating, exercise, social program take-up, organ donation, environmental conservation, and other behaviors thought to benefit individuals or society.

Behavioral scientists have offered two intuitive reasons why nudges might increase welfare. The first is cost effectiveness: nudges might increase ostensibly beneficial behaviors at low implementation cost (Benartzi et al. 2017; Loewenstein and Chater 2017). Correspondingly, many empirical evaluations focus on average treatment effects. The second is "asymmetric" or "libertarian" paternalism: nudges might improve choices by biased consumers without affecting those who fully optimize (Camerer et al. 2003; Thaler and Sunstein 2003). But to what extent do these intuitive arguments coincide with standard economic notions of total surplus maximization? And what parameters would we need to measure to quantify the welfare effects of nudges?

We address these questions with theoretical tools from public finance and novel randomized experiments evaluating automotive fuel economy labels and sugary drink health labels. Our theoretical results clarify that despite being the focus of much empirical work, *average* treatment effects are an incomplete and potentially misleading way to evaluate nudges. Measuring heterogeneity in treatment effects, including how they covary with mistakes or externalities, is crucial for welfare analysis. We show that the sufficient statistics for treatment effect heterogeneity are distinct from earlier notions of libertarian or asymmetric paternalism, particularly in markets where demand shocks affect producer prices.

In our model, heterogeneous consumers make a binary choice (for example, whether to buy a sugary drink or a sugar-free alternative) in a market with three distortions: heterogeneous consumer biases, heterogeneous consumption externalities, and producer market power. The government has two policy instruments: a uniform tax paid by producers, which may not be set optimally due to political or technical constraints, and a "nudge," which changes each consumer's relative willingness-to-pay (WTP) by a potentially heterogeneous amount τ. The heterogeneity of τ differentiates the nudge from the tax and is central to the welfare analysis.

We show that the nudge's total surplus effect depends on two key forces. First, with non-zero pass-through of producer taxes to prices, and a non-optimal tax, the nudge's total surplus depends on whether it has *beneficial average effects*: formally, if the average of WTP effects τ counteracts the average net distortion from bias, externalities, and market power that is not already internalized by a tax. This supports the common intuition that increasing ostensibly "beneficial" behaviors

increases welfare. However, average effects become less important at lower pass-through rates. In the limiting case with zero pass-through, the nudge cannot change the equilibrium quantity of the good provided by the market, and all that matters is the efficiency of the allocation of that fixed quantity. Analogously, if the government can optimally control aggregate quantity through an optimal tax, average effects are again irrelevant. Tangibly, this means that if governments can set optimal taxes on goods like gas guzzling cars, sugary drinks, or cigarettes, *it does not matter whether a nudge decreases or increases average consumption.*

The second key determinant of the nudge's total surplus effect is the extent to which it reduces the variance of choice distortions caused by bias and externalities. For example, hard information could reduce distortion variance if it has the largest effects on the most misinformed consumers. Alternatively, a graphic warning label could be adversely targeted if consumers who already paid more attention to a product's harms also pay more attention to the label. An imprecise label could also increase distortion variance if different consumers interpret it differently, adding idiosyncratic noise to choices. But, more subtly, hard information could also increase distortion variance if biases are relatively homogeneous and the information debiases only a fraction of consumers.

These results clarify how asymmetric paternalism, libertarian paternalism, and improvements in "deliberative competence" (Ambuehl, Bernheim and Lusardi 2022) translate to market settings and cases where the government can also set taxes. With endogenous producer prices and/or optimal taxes that adjust to nudges, nudges can reduce total surplus even if they improve the decisions of some biased consumers without affecting optimizing consumers. For example, a market with homogeneous bias can still achieve the first best if supply is perfectly inelastic or the government can set the optimal tax. In that case, it is heterogeneity in bias that generates misallocation, so nudges can reduce efficiency if they debias some consumers but not others.

We quantify the importance of these issues by applying the theory to two online randomized experiments. In the "cars experiment," consumers make binary choices between hypothetically leasing 23 mile-per-gallon (MPG) and 30 MPG sedans. In the "drinks experiment," consumers make binary choices between sugary and sugar-free drinks with similar flavors. The two experiments have parallel structures: we first measure behavioral bias, then measure baseline relative WTP using a multiple price list (MPL), and finally measure endline WTP while displaying fuel economy or health labels.

We chose labels that have been implemented or proposed by researchers or government agencies. In the cars experiment, we randomized four labels: (i) the MPG component of the fuel economy label required by the U.S. Environmental Protection Agency (EPA), (ii) the average annual fuel cost component of the EPA label, (iii) a personalized annual fuel cost label based on each participant's 2019 miles driven and gas price, and (iv) the EPA's "SmartWay" environmental label. In the drinks experiment, we randomized three labels: (i) a magnified nutrition facts label, (ii) a stop sign health warning label recommended by Grummon et al. (2019b), and (iii) the graphic warning label studied by Donnelly et al. (2018). The label type was randomly assigned across participants, and a control group saw no labels on their endline choices.

Both experiments were incentivized. In the cars experiment, participants' payments depended on how close their relative WTPs were to their "true" relative values implied by earlier survey responses. In the drinks experiment, we randomly selected just over 20 percent of participants and shipped them the drink they had chosen in one of their MPL choices, while deducting the price of the drink from their payment.

In the cars experiment, behavioral bias is the difference between baseline relative WTP and the "true" relative value. As in Ambuehl, Bernheim and Lusardi (2022), these are objective measures of decision quality within our experiment, because bias is defined as the difference between elicited WTP and the WTP that would have maximized the participant's experimental payment. Relative externalities depend on how much each participant drives, the incremental fuel use per mile for the 23-MPG versus 30-MPG car, and the social cost of carbon. In the drinks experiment, we estimate bias from survey measures of nutrition knowledge and self-control using the approach of Allcott, Lockwood and Taubinsky (2019a). We assume a homogeneous health system cost externality for sugary drinks, again following Allcott, Lockwood and Taubinsky (2019a).

The two experiments are complementary in how they trade off internal and external validity. In the cars experiment, we have an objective measure of behavioral bias within the experiment, but the decisions didn't involve actual car purchases. In the drinks experiment, we impose strong assumptions to estimate bias, but participants could actually receive the drinks they chose.

The within- and between-subject variation in labels allows us to measure both average treatment effects and heterogeneity. In the label treatment groups, the relative WTP changes between baseline and endline choices represent the person-specific effect τ plus elicitation noise. The control group WTP changes isolate elicitation noise. Thus, the average and variance of WTP changes for treatment relative to control identify the average and variance of τ. A label's effect on distortion variance depends on how much of the variance in τ represents a covariance with bias and externalities. Our econometric approach addresses potential measurement error in individual-level estimates of bias, externalities, and treatment effects.

In the cars experiment, we estimate small distortions toward the lower-MPG cars, with relative biases and externalities summing to an average of about 3 percent of the typical price. In the drinks experiment, we estimate large distortions toward sugary drinks, with relative biases and externalities summing to 95 percent of price. All fuel economy and sugary drink labels reduce demand for lower-MPG cars and sugary drinks. However, the labels' average effects are much smaller than the average distortion, consistent with the view that nudges are not sufficiently powerful to completely substitute for taxes (Loewenstein and Chater 2017; Thaler and Sunstein 2021).

We estimate that the labels increase the variance of distortions. The label effects have substantial variance: the estimated coefficients of variation (standard deviation of τ / absolute average τ) are 3.0 and 2.0 when pooling across labels in the cars and drinks experiments, respectively. Fuel economy labels add purely idiosyncratic noise to consumers' decisions: the label effects τ have statistically zero covariance with bias and externalities. Sugary drink labels (especially the stop sign and graphic warnings) are adversely targeted: they reduce WTP for sugary drinks more for

consumers estimated to have *less* bias.

The total surplus calculation delivers a striking result: although both fuel economy and health warning labels have beneficial average effects, they do not necessarily increase total surplus in our model. With no tax, hypothetical labels with homogeneous effects equal to the estimated average τ would substantially increase total surplus. However, the increased distortion variance erases much of the surplus gain from sugary drink labels and actually causes fuel economy labels to reduce total surplus. In sensitivity analyses, we find that plausible alternative values of the effects on distortion variance can change the sign of total surplus effects and have enormous implications for the magnitudes. Similarly, alternative pass-through values within the range of previous estimates also significantly affect the results by changing the relative importance of average effects.

When the government can also set the optimal tax, the average treatment effects are irrelevant, so both fuel economy and sugary drink labels reduce total surplus in our model. This is important because some U.S. cities now tax sugary drinks, and the U.S. uses corporate average fuel economy standards and gasoline taxes to increase demand for higher-MPG cars. If those policies are set close to optimally, our estimates suggest that adding labels might *reduce* total surplus in our model by adding noise to consumer choice.

As an example of how average effects can be misleading, compare sugary drink warning labels to the magnified nutrition facts label. Our point estimates suggest that graphic warning labels reduce average WTP for sugary drinks more than nutrition facts labels, so an evaluation based only on average effects might suggest that graphic warning labels are preferred. However, graphic warning labels cause larger increases in distortion variance than nutrition facts: their treatment effects have higher variance overall (possibly because graphic warnings have more ambiguous interpretations), and their effects are smaller for more biased consumers. Thus, our point estimates suggest that graphic warning labels deliver lower total surplus in our model, despite having larger average effects. Additional survey evidence suggests that graphic warning labels are highly aversive, which further worsens their total surplus effects.

A key limitation of our empirical work is that estimating individual-level variation in biases and nudge effects requires empirical designs that would have been especially challenging in naturalistic settings. We chose to use more artefactual designs to cleanly illustrate how to estimate the required sufficient statistics, but it is unlikely that our exact parameter estimates generalize outside our experiments. However, we suspect that our qualitative conclusion about the importance of distortion variance generalizes to many other settings.

Our paper builds on previous work in several areas. Our theoretical framework builds on Farhi and Gabaix (2020), who derive first-order conditions for socially optimal nudge intensity in a perfectly competitive market with constant marginal costs, under parametric assumptions about how nudges affect utility. We use techniques from Weyl and Fabinger (2013) to study the welfare effects of nudges with fewer parametric assumptions and in a more general setting with market power and nonlinear production costs. This makes our analysis applicable to the many markets with incomplete pass-through, including autos and sugary drinks (Sallee 2011; Allcott, Lockwood

and Taubinsky 2019b). We also go beyond Farhi and Gabaix (2020) by deriving welfare formulas based on empirically measurable sufficient statistics and then estimating those parameters in two applications.

Our empirical work extends a rich empirical literature that estimates the effects of nudges but stops short of welfare analysis.[1] Our cars experiment extends the empirical literature on durable good energy use information disclosure (e.g., Davis and Metcalf 2016; Allcott and Sweeney 2017; Allcott and Knittel 2019; Rodemeier and Loschel 2022). Our drinks experiment extends the empirical literature on sugary drink warning labels (e.g., Donnelly et al. 2018; Moran and Roberto 2018; Grummon et al. 2019a; Grummon et al. 2019b; Grummon and Hall 2020; Taillie et al. 2020; Hall et al. 2022) and nutrition information provision more broadly (e.g., Bollinger, Leslie and Sorensen 2011; Kiesel and Villas-Boas 2013; Araya et al. 2022; Barahona et al. 2022). While many of these papers estimate heterogeneous treatment effects, none estimate the full set of parameters that we show are required for welfare analysis. By clarifying the required parameters and showing how they can be estimated, we hope that our work provides a template for empirical welfare analysis.

Our paper also extends a smaller group of papers that quantitatively evaluate the welfare effects of nudges (e.g., Carroll et al. 2009; Handel 2013; Chetty et al. 2014; Bernheim, Fradkin and Popov 2015; Damgaard and Gravert 2018; Houde 2018a, 2018b; Allcott and Kessler 2019; Finkelstein and Notowidigdo 2019; Thunström 2019; Goldin and Reck 2020; Ambuehl, Bernheim and Lusardi 2022; Butera et al. 2022; Barahona, Otero and Otero 2023; List et al. 2023; Choukhmane and Palmer 2023).[2] Most broadly, our work extends the behavioral public economics literature studying optimal policy in the presence of behavioral bias; see Mullainathan, Schwartzstein and Congdon (2012) and Bernheim and Taubinsky (2018) for reviews. We differ from previous work in that we (i) provide a general public finance framework for evaluating corrective nudges that is portable across different markets and types of nudges, (ii) derive sufficient statistics for welfare evaluation, including the important but sometimes nuanced effects of treatment effect heterogeneity, and (iii) use two experiments to demonstrate how the required heterogeneity and targeting parameters could be estimated.[3] The goal of our paper is to provide a reasonably complete analysis of nudges in product markets, which covers many, but certainly not all, important economic settings where

[1] This includes evaluations of nudges in contexts such as charitable giving (Damgaard and Gravert 2016; Exley and Kessler 2019), education (Hastings and Weinstein 2008; Jensen 2010; Allende, Gallego and Neilson 2019), energy conservation (Schultz et al. 2007; Nolan et al. 2008; Allcott 2011, 2015; Allcott and Rogers 2014; Ito, Ida and Tanaka 2018; Knittel and Stolper 2019; Loschel, Rodemeier and Werthschulte 2022), retirement savings (Madrian and Shea 2001; Chetty et al. 2014; Beshears et al. 2015), smoking (Gine, Karlan and Zinman 2010; Thrasher et al. 2012; Hammond et al. 2012; Cantrell et al. 2013; Brewer et al. 2016; Kenkel et al. 2023), social program takeup (Bhargava and Manoli 2015), vaccinations (Milkman et al. 2021, 2022), and voting (Gerber and Rogers 2009).

[2] Choukhmane and Palmer (2023) apply our framework to natural experiments on retirement savings policy.

[3] The important independent work by List et al. (2023) is complementary, because their theory and empirical strategy cannot be applied to the settings we study. Their theory considers a special case of our model where the nudge offsets each consumer's bias by some homogeneous proportion. This would not fit our data, where nudges have heterogeneous effects that are not closely related to bias. The empirical strategy (using meta-analyses) applies to a case where the variance in average effects *across* different nudges and settings equals the variance of one nudge's effects across consumers *within* a market. This again would not fit our data, where the heterogeneity in average effects across different nudges is not related to the heterogeneity in one nudge's effects across consumers.

nudges are currently used.

Sections 1–5 present the theoretical framework, experimental designs, parameter estimation, welfare analysis, and conclusion, respectively. All proofs are in Appendix A.3.

1 Theoretical Framework

1.1 Setup

We model a unit mass of consumers who choose whether or not to buy a good that delivers utility v at price p. We assume quasilinear utility, so the utility gain from buying the good is $v - p$. Consumers overestimate their utility from the good by amount γ, due to forces such as inattention, projection bias, and imperfect information. We refer to γ as consumer "bias."

The government has two instruments: a linear tax t on producers and a nudge, both of which have zero implementation cost. Because the tax is on producers, it affects consumer demand only through producer prices, so it is not heterogeneously perceived due to salience effects, as in Chetty, Looney and Kroft (2009), Taubinsky and Rees-Jones (2018), and others. The nudge increases willingness-to-pay by amount $\sigma\tau$, where $\sigma \in \mathbb{R}^+$ is the nudge intensity and τ covaries arbitrarily with v and γ. Accounting for the effects of bias and the nudge, consumers buy the good if $v + \gamma + \sigma\tau > p$.

v, γ, and τ vary across consumers. We assume that the joint distribution $F(v, \gamma, \tau)$ generates smooth demand curves, but make no other assumptions. Allowing τ to be a function of v and γ would be mathematically equivalent to imposing a particular covariance between τ and v and γ. We define $D(p, \sigma)$ as the demand curve, D'_p as its derivative with respect to p, and $\varepsilon_D = -pD'_p/D$ as its elasticity.

The nudge may also impact consumption utility or impose a direct psychic cost or benefit. For example, cigarette graphic warning labels might make smokers feel worse, while labels promoting healthy foods might make consumers feel better. We define $\sigma\iota_1$ and $\sigma\iota_0$ to be these direct effects on consumers who buy and don't buy the good, respectively. We define $I(\sigma, p)$ as aggregate psychic benefits, which may indirectly depend on p if $\iota_1 \neq \iota_0$ for some consumers.

On the supply side, we follow the Weyl and Fabinger (2013) model of symmetric competition. We normalize such that there is also a unit mass of firms, and we let firm j's cost of producing quantity q_j be $c(q_j)$. We limit to symmetric equilibria where each firm produces $q_j = q$, giving aggregate quantity q, market price $p(q)$, and markup $\mu \coloneqq p - c'(q) - t$. Weyl and Fabinger (2013) show that a wide range of firm conduct models can be captured by an elasticity-adjusted Lerner index $\theta := \frac{\mu}{p}\varepsilon_D$ that is constant. For example, homogeneous-product Bertrand competition has $\theta = 0$, Cournot competition with n firms has $\theta = 1/n$, monopoly has $\theta = 1$, and the Delipalla and Keen (1992) conjectures model also has constant θ. We define $\rho := \frac{dp}{dt}$ as the pass-through rate.

Total surplus W is the sum of consumer surplus, producer surplus, and government revenue. In an efficient market, all consumers with $v > c'(q)$ would buy the good. We refer to $(\gamma + \sigma\tau - \mu)$ as the "net distortion" to consumer choice caused by bias, the nudge, and the markup. We ignore

externalities in this section for simplicity, but without loss of generality. If consumption generates a negative externality ϕ, then the net distortion and total surplus formulas hold after replacing γ with $(\gamma + \phi)$.

In our empirical applications, consumers choose between two goods. This framework applies directly if we redefine v, p, γ, and $\sigma\tau$ as the *relative* utility, price, bias, and nudge effect for the first good compared to the second, and let $D(p)$ be the demand for the first good. We consider the general case of many goods with endogenous prices in Appendix A.4.

For any function $X(v, \gamma, \sigma, \tau)$ we define $\mathbb{E}_m[X(v, \gamma, \sigma, \tau)]$ to be the conditional expectation of X over the set of marginal consumers, i.e., the set $\{(v, \gamma, \tau)|v + \gamma + \sigma\tau = p\}$. With a slight abuse of notation, we define

$$\frac{\partial}{\partial \sigma}\mathbb{E}_m[X(v, \gamma, \sigma, \tau)] := \frac{d}{d\sigma'}\int_{\{(v,\gamma,\tau)|v=p-\gamma-\sigma\tau\}} X(v, \gamma, \sigma', \tau)dF\,|_{\sigma'=\sigma}\,. \tag{1}$$

for any function X. The partial derivative notation symbolizes that we take the derivative of the function X, but not of the set of consumers over which the expectation is taken. We use analogous notation for variance and covariance operators.

1.2 Motivating Examples

Before presenting formal results about the welfare effects of nudges, we use stylized examples to demonstrate the core intuitions. To simplify the examples, we consider the effects of changing σ from 0 to 1. We also assume for the examples that v is distributed uniformly and independently of γ, and that for each tuple $(\gamma, \sigma, \tau, t)$ there is always a consumer on the margin.[4] We abstract away from psychic costs, so that a nudge that exactly offsets the bias (i.e., $\tau = -\gamma$ for all consumers) delivers the social optimum in the absence of market power or taxes. All conclusions in these examples can be derived formally from Proposition 1.

For the first five examples, we assume a competitive market. For the first three, we also assume constant marginal cost.

Example 1. Consider a sugary drink market with two types of consumers. "Oblivious" consumers are inattentive to health harms from sugary drinks ($\gamma_o > 0$) and don't pay attention to health warning labels ($\tau_o = 0$). "Health-conscious" consumers are initially unbiased ($\gamma_h = 0$), but warning labels make them over-sensitized to health ($\tau_h < 0$). On average, consumers over-consume sugary drinks ($\mathbb{E}[\gamma] > 0$), and the label reduces consumption ($\mathbb{E}[\tau] < 0$). Thus, a naive analysis might conclude that the labels increase welfare because they have "beneficial" average effects. In reality, however, the label reduces total surplus because it is poorly targeted: it distorts the choices of unbiased consumers without improving the choices of biased consumers.

This example is plausible: psychological theories such as those of Ungemach et al. (2017) predict that product labels can serve as "signposts" that activate "dormant objectives" (Bond, Carlson

[4] In other words, we assume that there is a value of v in the support of the distribution such that $v + \gamma + \sigma\tau = p$.

and Keeney 2008) such as health-consciousness or environmentalism, potentially causing oversensitivity to these attributes. Value activation could cause poor targeting if health-conscious or environmentalist consumers are less biased because they are already aware of a product's nutritional or environmental characteristics.

Example 2. Now suppose there are two different types of consumers. "Overconsumers" (2/3 of the population) have $\gamma_o > 0$, and a nudge offsets half of their bias ($\tau_o = -\gamma_o/2$). "Underconsumers" (1/3 of the population) have bias of the same magnitude but in the opposite direction ($\gamma_u = -\gamma_o < 0$), and the nudge fully offsets their bias ($\tau_u = -\gamma_u$). On average, consumers overconsume the good ($\mathbb{E}[\gamma] > 0$), but the nudge does not change consumption ($\mathbb{E}[\tau] = 0$). Thus, a naive analysis might conclude that the nudge doesn't increase welfare because it has zero average effect. In reality, however, the nudge increases total surplus because it is well-targeted to offset both positive and negative biases.

This example might be relevant for some types of information disclosure: consumers have wide dispersion in beliefs about product characteristics such as calorie content and energy use, and hard information is designed to reduce this dispersion (Bollinger, Leslie and Sorensen 2011; Allcott and Taubinsky 2015).

Example 3. The previous two examples illustrate how a negative covariance between nudge effects τ and bias γ can be key to increasing total surplus. However, this does not guarantee that a nudge increases total surplus. Suppose that the nudge effects are now $\tau = -\gamma + \varepsilon$, where ε is mean-zero noise that is independent of γ. Thus, the covariance between τ and γ is negative and $\mathbb{E}[\tau|\gamma] = -\gamma$, meaning that on average, the nudge fully offsets the bias. Without the nudge, consumers buy the good if $v + \gamma \geq p$, while with the nudge, consumers buy if $v + \varepsilon \geq p$. Thus, the nudge eliminates one distortion (γ) but adds another (ε), and it will decrease allocative efficiency if ε has sufficiently high variance.

This example is relevant in settings where consumers might interpret information differently, or when information is not personalized. For example, energy cost labels on appliances and cars in the U.S. report energy costs at national average utilization and energy prices. This averaging adds noise relative to each consumer's personalized values, as studied by Davis and Metcalf (2016).

Example 4. Suppose that consumers have a homogeneous bias γ. A nudge fully debiases half of consumers ($\tau_1 = -\gamma$) but does not affect that other half ($\tau_2 = 0$). Supply is fixed at some value $q^\dagger$, which implies pass-through $\rho = 0$, since the equilibrium price must always equal the value at which demand meets the fixed supply. Without the nudge, the market is efficient, despite the consumer bias. This is because the share of consumers buying the good always equals $q^\dagger$, and when bias is homogeneous, the consumers buying the good will always be the share $q^\dagger$ of consumers with the highest v. However, with the nudge, the consumers buying the good will be the $q^\dagger$ consumers with the highest $v + \gamma + \tau$, which will not be the $q^\dagger$ consumers with the highest v when $\gamma + \tau$ is heterogeneous. Thus, the nudge reduces total surplus by increasing the variance of the net distortion.

While few markets have fully inelastic supply, we will see that this intuition applies to any market with $\rho < 1$, which is empirically very common.

Example 5. As in the previous example, suppose that consumers have homogeneous γ and the nudge fully debiases half of consumers without affecting the other half. Without the nudge, a Pigouvian tax of $t = \gamma$ can achieve the first best. With the nudge, however, the net distortion $\gamma + \tau$ is heterogeneous, so a uniform tax cannot achieve the first best whenever $\rho > 0$. Thus, the nudge again reduces total surplus by increasing the variance of the net distortion. However, if the tax is constrained to $t = 0$, the nudge increases total surplus for a value of ρ sufficiently close to 1. This illustrates how taxes and nudges can be substitutes. Alternatively, the tax and nudge could be complements if nudges make net distortions more homogeneous.

This example is crucial because policymakers control both taxes and nudges in many markets.

Example 6. Suppose that there is no tax, and producers have market power, giving markup $\mu > 0$. Any bias less than the markup ($0 < \gamma \leq \mu$) reduces the net distortion—the two market failures offset. Thus, any nudge that offsets bias ($-\gamma \leq \tau < 0$) reduces total surplus. At any tax t, the first best obtains if there is homogeneous bias of $\gamma = \mu + t$, so a nudge could only reduce total surplus.

These examples illustrate that the welfare effects of nudges depend on many factors other than whether they have beneficial average effects. This underscores the importance of a full characterization that accounts for heterogeneity, pass-through, market power, and taxes, which we present below.

1.3 Effects of Nudges on Total Surplus

1.3.1 Graphical Illustration

Figure 1 presents a graphical illustration of the effect of a nudge on total surplus. For the figure (only), we assume full pass-through ($\rho = 1$), zero psychic cost ($I = 0$), homogeneous bias γ and nudge effect τ, and a linear demand curve. The solid lines illustrate market demand and supply; their intersection gives the initial equilibrium at point F. The dashed lines illustrate aggregate marginal utility and aggregate marginal cost; their intersection gives the social optimum at point A. The vertical distance between market demand and welfare-relevant demand is the bias γ. The vertical distance between price and aggregate marginal cost is the markup plus tax $\mu+t$. A marginal nudge $d\sigma$ shifts demand by $\tau \cdot d\sigma$; the demand decrease (as drawn) corresponds to $\tau < 0$.

The deadweight loss at the initial equilibrium relative to the social optimum is the triangle $ABC = -\frac{1}{2}\{\gamma + \sigma\tau - \mu - t\}^2 D'_p$. The deadweight loss after a small increase $d\sigma$ is $AED = -\frac{1}{2}\{\gamma + (\sigma - d\sigma)\tau - \mu - t\}^2 D'_p$. Equivalently, $AED = ABC - \frac{1}{2}(d\sigma)\frac{\partial}{\partial\sigma}\{\gamma + \sigma\tau - \mu - t\}^2 D'_p + O((d\sigma)^2)$. Thus, when $d\sigma$ is small, so that terms of order $(d\sigma)^2$ are negligible, the total surplus gain is given by $BCDE \approx \frac{1}{2}(d\sigma)\frac{\partial}{\partial\sigma}\{\gamma + \sigma\tau - \mu - t\}^2 D'_p$.

Now returning to the case with heterogeneous $\{\gamma, \tau\}$, the change in total surplus from the marginal nudge increase $d\sigma$ is the average area of $BCDE$ areas over all marginal consumers:

$$dW = \frac{1}{2}(d\sigma)\frac{\partial}{\partial\sigma}\mathbb{E}_m\left[(\gamma + \sigma\tau - t - \mu)^2\right] D'_p. \tag{2}$$

1.3.2 Formal Result

Proposition 1 formally presents the effects of a marginal nudge intensity increase in two cases: exogenous taxes (perhaps constrained by technical or political factors) and optimal taxes.

Proposition 1. *Assume that $\frac{d}{dp}\varepsilon_D$ and $\frac{d}{d\sigma}\varepsilon_D$ are negligible whenever $\mu > 0$. At a fixed tax t, the total surplus change from a marginal increase in nudge intensity is*

$$\frac{dW}{d\sigma} = \frac{1}{2}\frac{\partial}{\partial\sigma}\left\{\rho\underbrace{\left(\mathbb{E}_m\left[\gamma + \sigma\tau - t - \mu\right]\right)^2}_{\text{average distortion}} + \underbrace{Var_m\left[\gamma + \sigma\tau\right]}_{\text{distortion variance}}\right\} D'_p + \frac{\partial I}{\partial\sigma} + (1-\rho)\mathbb{E}_m[\tau]\frac{\partial I}{\partial p}. \tag{3}$$

At the optimal tax $t^* = \mathbb{E}_m\left[\gamma + \sigma\tau\right] - \mu - \sigma\mathbb{E}_m\left[\iota_1 - \iota_0\right]$, the total surplus change from a marginal increase in nudge intensity is

$$\frac{dW}{d\sigma} = \frac{1}{2}\frac{d}{d\sigma}Var_m\left[\gamma + \sigma\tau\right] D'_p + \frac{\partial I}{\partial\sigma} + \mathbb{E}_m[\tau]\frac{\partial I}{\partial p}. \tag{4}$$

For intuition, consider the case with non-optimal tax t, full pass-through ($\rho = 1$), and zero psychic cost ($I = 0$). This makes equation (3) match the graphical illustration and equation (2), after expanding the second moment. Further rearranging gives

$$\frac{dW}{d\sigma} = \frac{1}{2}\frac{\partial}{\partial\sigma}\left\{\underbrace{\left(\mathbb{E}_m\left[\gamma\right] - \mu - t + \sigma\mathbb{E}_m\left[\tau\right]\right)^2}_{\text{average distortion}} + \underbrace{2\sigma Cov_m\left[\gamma,\tau\right] + \sigma^2 Var_m\left[\tau\right]}_{\text{distortion variance}}\right\}\underbrace{D'_p}_{<0}. \tag{5}$$

Since $D'_p < 0$, more negative values of the terms inside brackets imply more positive total surplus effects.

The first term inside brackets, $\left(\mathbb{E}_m\left[\gamma\right] - \mu - t + \sigma\mathbb{E}_m\left[\tau\right]\right)^2$, captures the common intuition that nudges increase total surplus if they have *beneficial average effects*: formally, if $\mathbb{E}_m\left[\tau\right]$ has the opposite sign of the distortion from biases and the markup that is not already internalized by the tax. The remaining terms inside brackets capture the extent to which the nudge reduces the variance of choice distortions caused by bias and externalities. This is the sum of two forces. First, holding constant the variance of nudge effects $Var_m\left[\tau\right]$, how much do the nudge effects τ covary with bias γ? The term $Cov_m\left[\gamma,\tau\right]$ formalizes Examples 1 and 2, which illustrated how nudges increase surplus when treatment effects covary negatively with bias. Second, holding constant $Cov_m\left[\gamma,\tau\right]$, how large is the variance in τ? The term $Var_m\left[\tau\right]$ formalizes Example 3, which illustrated how surplus decreases when nudges add idiosyncratic noise to choices that is unrelated to distortions.

Now consider a different special case of equation (3) with zero pass-through ($\rho = 0$), corresponding to fixed quantity supplied. The equation becomes identical to equation (4), where the average effect is irrelevant and the total surplus effect depends only on $Var_m[\gamma + \sigma\tau]$. With fixed supply, the only way for a nudge to improve allocative efficiency is to reduce the variance of the net distortion, as Example 4 illustrated. Equation (3) shows that the importance of the average effect increases linearly with pass-through ρ.

Next, consider equation (4), the effect of a nudge when the government also sets an optimal tax. The intuition is similar to the previous case: when the government controls aggregate quantity through an optimal tax, the only way for a nudge to improve allocative efficiency is again to reduce the variance, as Example 5 illustrated.

The net distortion also depends on the markup μ, as Example 6 illustrated. When a positive bias offsets distortions from market power, using a nudge to offset the bias is inefficient. Conversely, when a negative bias depresses demand, this is especially harmful in the presence of market power, and nudges that offset the bias are especially beneficial. This is reflected in the term $\mathbb{E}_m\left[(\gamma + \sigma\tau - \mu - t)^2\right]$ in equation (3). Because the optimal tax is set equal to the average marginal net distortion (including μ), μ does not appear in equation (4).

Proposition 1 shows that the nudge has two effects on psychic benefits. $\frac{\partial I}{\partial \sigma}$ is the direct effect on psychic benefits holding prices constant. $(1-\rho)\mathbb{E}_m[\tau]\frac{\partial I}{\partial p}$ is an indirect effect: the nudge affects prices, which in turn affect how many consumers incur the psychic benefits associated with purchasing or not purchasing the product, i.e., $\sigma\iota_1$ versus $\sigma\iota_2$. When taxes are endogenous, the nudge affects price in two ways: through its effect on prices holding the tax constant, and through its effect on the optimal tax. Because of this, the factor multiplying $\frac{\partial I}{\partial p}$ given an optimal tax is higher.

1.4 Asymmetric Paternalism, Libertarian Paternalism, and Deliberative Competence in Markets

Our results clarify how asymmetric paternalism, libertarian paternalism, and improvements in "deliberative competence" (Ambuehl, Bernheim and Lusardi 2022) translate to market settings. Camerer et al. (2003, page 1212) write that "a regulation is asymmetrically paternalistic if it creates large benefits for those who make errors, while imposing little or no harm on those who are fully rational." Thaler and Sunstein (2003, page 175) write that paternalism is "libertarian" "if no coercion is involved," and Thaler and Sunstein (2008, page 6) write that "a nudge ... is any aspect of choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives."

Except in perfectly competitive markets with fully elastic supply, nudges indirectly change equilibrium prices. In Appendix A.2, we show that the effect of a marginal nudge increase on price is

$$\frac{dp}{d\sigma} = (1-\rho)\mathbb{E}_m[\tau]. \tag{6}$$

Nudge-induced price changes harm or benefit all consumers, including those who are fully rational, potentially violating the definition of libertarian or asymmetric paternalism.[5] Furthermore, this shows how nudges indirectly affect economic incentives, which is not contemplated in the definition of a "nudge." Note that the price change is not "insignificant": equation (6) shows that unless $\rho = 1$, nudges have first-order effects on prices that scale with the average marginal τ.

A natural modification might be to define an instrument as "asymmetrically paternalistic" if it weakly improves decision quality for all consumers without affecting decisions by unbiased consumers: $|\gamma + \tau| \leq |\gamma|, \forall (\gamma, \tau)$. Relatedly, Ambuehl, Bernheim and Lusardi (2022) say that a nudge improves "deliberative competence" if $\mathbb{E}\left[(\gamma + \sigma\tau)^2\right] < \mathbb{E}\left[\gamma^2\right]$ in our notation. (See Appendix A.1 for more in-depth discussion.) Proposition 1.3 and the preceding examples show that these concepts do not guarantee total surplus gains outside of the special case of no taxation, full pass-through, no markup, and no psychic costs.

2 Experimental Design and Data

Our empirical analyses focus on two markets (automobiles and packaged drinks) that both involve externalities, possible consumer bias, and various types of information labels. There are many models of automobiles and types of packaged drinks, but we simplify the design and analysis to focus on binary choices between closely comparable low- versus high-fuel economy sedans and sugary versus sugar-free drinks.

For cars, the potential biases we study are imperfect information about or inattention to gasoline costs, following Busse, Knittel and Zettelmeyer (2013), Allcott and Wozny (2014), Allcott and Taubinsky (2015), Sallee, West and Fan (2016), Allcott and Knittel (2019), Gillingham, Houde and van Benthem (2019) and others. For sugary drinks, the potential biases we study are self-control problems and imperfect information about health costs, following Allcott, Lockwood and Taubinsky (2019a) and others.

We use two randomized experiments to identify utility v, bias γ, and treatment effects τ. The two experiments share a common format:

1. introductory questions,
2. a baseline multiple price list (MPL) that elicits relative valuations of the two choices, and
3. an endline MPL with information labels.

Screenshots from the two experiments are available in Appendix B.

2.1 Cars Experiment

Sample source. We fielded the cars experiment in April 2021 on Amerispeak, an online panel managed by the National Opinion Research Corporation (NORC) at the University of Chicago.

[5] Appendix A.2 provides formulas for the effect of nudges on consumer and producer surplus.

AmeriSpeak is a probability-based panel designed to be representative of the U.S. population. NORC randomly selects U.S. households and recruits them by mail, telephone, and in-person visits in an effort to minimize refusals. To minimize fatigue, panelists take only an average of two to three surveys per month. This design makes AmeriSpeak more plausibly representative and higher quality than typical survey panels that allow anyone to opt in and take as many surveys as they like.

Introductory questions. The experiment began by asking people whether they currently have a car; people who did not were screened out. Participants were then asked their state of residence, the number of miles they (and anyone else in their household) drove their primary car in 2019, and the average price they paid for gas in 2019. We asked about 2019 because we wanted to ignore disruptions caused by the coronavirus pandemic. To ensure accurate responses, participants were asked to enter their annual mileage twice, and we asked people to confirm their gas price if their reported price differed from their state's 2019 average by more than $0.50 per gallon.

Participants were told,

For the entire survey, imagine the following scenario:

- *Your primary car broke down and can't be used anymore.*
- *To replace your primary car, you will lease a car for the next 3 years.*
- *Any dealership would offer you the same lease contract, so there's no need to shop around.*
- *You (and anyone else in your household) would drive this car the amount that you told us earlier.*

The experiment elicited yearly WTP to lease four mid-size sedans: the Honda Accord, Nissan Altima, Subaru Legacy, and Ford Fusion. The Accord, Altima, and Legacy each get 30 miles per gallon (MPG), while the Fusion gets 23 MPG.

In the final introductory questions, participants were told, "We'd like to learn how much you would like to have each of the 4 cars on the next 4 screens, setting aside gas costs. In order to tell us that, please **imagine for the next 4 screens that gas is free** and that you'd still drive the [*2019 miles driven*] miles per year you told us earlier." Participants were shown the four cars and asked to write the maximum amount they'd pay per year to lease each one. In this elicitation and again in the MPLs described below, participants were shown each car's picture and key characteristics (including fuel economy), along with a link to more information on the manufacturer's website.

Baseline MPLs. Next, participants were told, "For the rest of the survey, assume that **gas is NOT free**, and the average price for the next three years will be the same as what you told us you paid in 2019." Using two baseline MPLs, we elicited people's relative WTP for the Altima vs. the Fusion and for either the Accord or Legacy vs. the Fusion. We used an adaptive MPL design that offered a series of binary choices between two cars at different prices. We randomized (i) whether the Altima-Fusion or Accord/Legacy-Fusion MPL was first, (ii) which of the Accord/Legacy appeared, and (iii) which car appeared on the left or right for each MPL.

In the adaptive MPL, the price of the car on the right was always \$2000 per year. The price of the car on the left began at \$2000 per year and would adjust for each subsequent question, reaching as low as \$500 or as high as \$3500, until a person's relative WTP was determined within \$100. People who were willing to pay less than \$500 or more than \$3500 for the car on the left were asked to report their maximum or minimum WTP in an open-answer box.

Endline MPLs with information labels. The endline MPLs were identical to the baseline MPLs, with two exceptions. First, we exchanged the Accord and Legacy, so people who had valued one at baseline now valued the other. Second, participants were randomly assigned to see fuel economy information labels just below the cars' pictures. Figure A3 presents an example of an endline binary choice screen, including each car's picture, a fuel economy label, and key characteristics.

There were five randomly assigned treatment conditions: full MPG label, average cost label, personalized fuel cost label, SmartWay label, and no label (control). Panel (a) of Figure 2 presents the four labels.

The full MPG and average cost labels were taken from the U.S. Environmental Protection Agency's official "Fuel Economy and Environment" label. In these treatment conditions, participants were told that the labels "are part of the labels that must be posted on car windows at dealerships" and that they "report averages for each car at typical usage patterns." The EPA's annual fuel costs are \$1850 for the 23-MPG Fusion and \$1450 for the 30-MPG Accord, Altima, and Legacy, assuming 15,000 miles driven per year and a gas price of \$2.87 per gallon.

The personalized cost label has the same format as the average cost label, but the numbers were calculated using each respondent's 2019 miles driven and gas price. In this condition, the participants were told that "the fuel cost numbers for each car are based on the annual miles driven and gasoline prices that you told us earlier."

The SmartWay label is an official EPA environmental certification given to vehicles that have relatively low emissions of greenhouse gases and local air pollutants. In this condition, participants were told that "Cars that meet SmartWay Vehicle standards for fuel economy and environmental performance will have SmartWay Vehicle labels. Cars that do not meet these standards will not have SmartWay Vehicle labels." The three 30-MPG cars are all SmartWay vehicles, while the 23-MPG Fusion is not.

Incentivization. Before the MPLs, participants were told that their responses would be incentivized. Specifically, participants were told that they would be paid an amount proportional to the "value" (i.e., consumer surplus) they would receive from an upcoming MPL choice. To compute this consumer surplus, we define f as the WTP if gas is free and g as the annual gas cost implied by the participant's 2019 miles driven and gas price and the vehicle's fuel economy rating. The "true" value implied by those previous survey responses is

$$\hat{v} = f - g, \tag{7}$$

and the consumer surplus at price p is $\hat{v} - p$.

The adaptive MPLs did not elicit choices at all possible price levels. We used the participant's responses to fill out the full "background" MPL that has one row for each possible price level. We then randomly selected one row from all baseline and endline background MPLs and computed the consumer surplus from the participant's choice in that row. Participants were told that it was in their best interest to answer all questions honestly, and were given a link to a page that contained the full instructions of how the adaptive MPL was used to fill out the "background" MPL. We rescaled that consumer surplus to be between \$1 and \$10, where \$1 and \$10 were set to the minimum and maximum possible consumer surplus achievable across all rows of all MPLs. We paid participants in dollar-equivalent "AmeriPoints" through the AmeriSpeak system.

Final sample. Out of 2,595 people who began the survey, 2,089 qualified for inclusion by reporting that they currently owned a car and successfully completing the survey. To increase the proportion of high-quality responses and improve precision in treatment effect variance estimates, we additionally dropped the 822 participants in the top or bottom five percent of annual gas cost, WTP if gas is free, estimated bias, or baseline-endline WTP change. Excluding these outlying responses is conservative: if we do not drop these 822 participants, the $Var[\tau]$ point estimate becomes much larger and $Cov[\phi, \tau]$ becomes more positive, strengthening the qualitative conclusion that fuel economy labels reduce welfare. The final sample size is 1,267 participants.

Appendix Table A1 presents descriptive statistics. To maximize precision, we do not weight the sample in our analyses. The unweighted sample has roughly similar demographics to the US population, although the sample has a higher share of people with a college degree. The average participant reported driving about 10,800 miles and paying \$2.79 per gallon for gas in 2019, both of which are very similar to the national averages. The average participant reported being willing to pay \$2,770 per year to lease one of the four cars, if gas were free, and the average relative WTP for the lower-MPG car from the baseline MPLs is about \$1,550 per year. Appendix Table A3 shows that the treatment groups are balanced on observables, except that the personalized fuel cost group is slightly younger and less white than the control group.

2.2 Drinks Experiment

Sample source. We fielded the experiment in fall 2021, recruiting participants through Facebook ads. See Appendix Figure A4 for an example ad. We recruited participants ourselves because we needed to ship drinks to participants' home addresses, and most online panels don't allow this or would charge prohibitively high prices.

Introductory questions. The experiment was run in two stages separated by three days. The first stage elicited contact information, demographic information, and bias proxies for nutrition knowledge and self-control. The bias proxy survey questions and variable construction are identical to Allcott, Lockwood and Taubinsky (2019a). Nutrition knowledge was measured with 28 questions from the General Nutrition Knowledge Questionnaire (GNKQ), which is widely used in the public health literature.[6] The nutrition knowledge bias proxy variable is the share of the 28 questions that

[6]One example question is, "If a person wanted to buy a yogurt at the supermarket, which would have the least

the participant answered correctly. Kliemann et al. (2016) show that the GNKQ has high construct validity, sensitivity to change, and test-retest reliability. Self-control was measured by people's level of agreement with the statement, "I drink soda pop or other sugar-sweetened beverages more often than I should." There were four responses: "Definitely," "Mostly," "Somewhat," and "Not at all." To construct the self-control bias proxy variable, we code those responses as 0, 1/3, 2/3, and 1, respectively. Allcott, Lockwood and Taubinsky (2019a) document that this question has high inter-rater reliability: people's ratings of their own self-control line up well with their spouse's ratings of them.

Baseline MPLs. Three days later, we sent emails inviting participants to complete the second stage. We included this three-day delay after the bias proxy questions because we did not want participants to have nutrition and self-control problems at top of mind as we elicited their valuations of sugary drinks. The second stage began by asking people to rate six sugary drinks from most favorite to least favorite: Minute Maid Lemonade, Coca-Cola, Pepsi, Seagram's Ginger Ale, Sprite, and Crush. Participants were told that they would be offered a series of choices between their three most favorite drinks and three substitutes.

Using three baseline MPLs, we elicited WTP for each participant's three favorite drinks relative to a sugar-free substitute with similar flavor.[7] As in the cars experiment, we used an adaptive MPL design that was still incentive-compatible because it was used to fill out a background MPL. We randomized the order of the three MPLs. Each choice screen showed the pictures of the drinks and a link to the nutrition facts.

In the adaptive MPL, the price of the drink on the right was always \$4, which is close to the typical market price of a 12-pack. The price of the drink on the left began at \$4 and would adjust for each subsequent question, reaching as low as \$1.20 or as high as \$6.80, until a person's relative WTP was determined within \$0.40. People who were willing to pay less than \$1.20 or more than \$6.80 were asked to report their maximum or minimum WTP in an open-answer box.

Endline MPLs with information labels. The endline MPLs were identical to the baseline MPLs, except that participants were randomly assigned to see magnified labels next to the drinks' pictures. There were four randomly assigned treatment conditions: nutrition facts label, stop sign warning label, graphic warning label, and no label (control). Panel (b) of Figure 2 presents the three labels.

In the nutrition facts label condition, participants were told that "we will add the official nutrition facts label to each drink." In the stop sign and graphic warning label conditions, participants were told, "Now, we will add nutritional warning labels to drinks with added sugar. Drinks without

sugar/sweetener?" The four possible responses were "0% fat cherry yogurt," "Plain yogurt," "Creamy fruit yogurt," and "Not sure." A second example is, "Which is the main type of fat present in each of these foods?" The five possible responses were "Polyunsaturated fat," "Monounsaturated fat," "Saturated fat," "Cholesterol," and "Not sure." This question was asked about olive oil (correct answer: monounsaturated), butter (saturated), sunflower oil (polyunsaturated), and eggs (cholesterol).

[7]The pairings of sugary drinks and sugar-free alternatives were: Minute Maid Lemonade and LaCroix Lemon, Coca-Cola and LaCroix Cola, Pepsi and LaCroix Cola, Seagram's Ginger Ale and 365 Ginger Sparkling Water, Sprite and LaCroix Lime, and Crush Orange and Bubly Orange.

added sugar will not have warning labels." The labels state, "WARNING: Drinking beverages with added sugar contributes to obesity, diabetes, and tooth decay." Both warning labels are taken from prior studies that found that they reduced demand for sugary drinks (Donnelly et al. 2018; Grummon et al. 2019a). Grummon et al. (2019b) recommend the version of the stop sign label we use because it appears to have larger effects on demand than other designs.

Finally, we included several questions to measure the labels' direct psychic benefits or costs. Participants were asked, "Imagine that drink manufacturers could print the above label on some drink containers. Also imagine that the computer selects you to receive the 12 pack of [a randomly chosen sugary drink]. Would you prefer to receive drink containers with that label or without it?" The survey then asked participants why they did or did not want the labels and used an MPL to elicit WTP to receive or avoid the labels.

Incentivization. Participants were told that they would receive a $15 completion payment via online gift card, and that their responses would be incentivized. Specifically, participants were told, "20 percent of respondents will be randomly selected to receive a 12-pack of drinks they chose in one of the survey questions plus an online gift card for the difference between $15 and the price of the drinks. The remaining respondents will receive their entire $15 reward in the form of an online gift card. For those who are selected to receive drinks, we will send the drinks to your address directly from either Walmart or Amazon." We initially selected one-third of participants for incentivization but later reduced the proportion due to logistical difficulties. Overall, we shipped drinks to 22 percent of participants.

Final sample. Out of the 5,744 people who consented to participate, 3,653 completed the first stage (including passing the attention check) and were invited to take the second survey. Of those, 2,619 completed the second stage (including passing the second attention check); this is our final sample size.

Table A2 presents descriptive statistics. The sample is wealthier and better educated than the US population. The average participant had a nutrition knowledge score of 0.70 and a self-control rating of 0.41. This latter number is lower than the US population average reported in Allcott, Lockwood and Taubinsky (2019a), perhaps because our recruitment ads were designed to recruit soda drinkers who would want to receive drinks as part of the experiment. Appendix Table A3 shows that the treatment groups are balanced on observables, except that the stop sign label group is slightly younger than the control group.

3 Experimental Results

This section presents reduced-form analyses of our two experiments and calculates sufficient statistics used in the welfare analysis in Section 4. We index consumers by i, product pairs (e.g., Altima & Fusion, or Coke & LaCroix Cola) by j, and choice occasions (baseline or endline MPLs) by $s \in \{1, 2\}$. All relative values between two goods (valuation v, bias γ, etc.) are for the more harmful relative to the less harmful good—i.e., the 23-MPG car minus the 30-MPG car, and the

sugary drink minus the zero-calorie drink.

We estimate population-level parameters instead of limiting to marginal consumers. In addition to the precision gains, this may increase generalizability, as market prices (and thus marginal consumers) vary over time and place. The parameter estimates are similar (although not identical) when we limit the sample to observations with WTP closer to the market prices at the time of the experiment; see Appendix Table A5.

3.1 Bias and Externalities

Both experiments deliver bias and externality estimates, which we denote as $\hat{\gamma}_{ij}$ and $\hat{\phi}_{ij}$, respectively. We assume that these are noisy measures of the true γ_{ij} and ϕ_{ij}, with mean-zero measurement error.

3.1.1 Cars Experiment

In the cars experiment, we estimate biases from imperfect information about or inattention to gasoline costs. A key advantage of our design is that we have an objective measure of bias within the experiment: bias is the extent to which people fail to maximize their expected experimental incentive by valuing the lower-MPG too high or too low on the MPLs. We define w_{ij1} as the baseline relative WTP for the lower-MPG car in product pair j, and $\hat{v}_{ij}$ is the relative "true" value defined in equation (7): WTP if gas is free minus gas costs. The bias estimate is

$$\hat{\gamma}_{ij} = w_{ij1} - \hat{v}_{ij}. \tag{8}$$

$\hat{\gamma}_{ij} = 0$ corresponds to a relative WTP that maximizes the hypothetical consumer surplus from leasing a car; this also corresponds to MPL choices that maximize the actual experimental incentive. $\gamma_{ij} > 0$ reflects a relative WTP above the incentive-maximizing level, while $\hat{\gamma}_{ij} < 0$ reflects a relative WTP that is too low.

Our externality estimate is the annual climate change damages from the additional gasoline consumption from driving the 23-MPG car instead of the 30-MPG car, based on each participant's 2019 miles driven.[8] We assume a $51 social cost of carbon (SCC), following the U.S. Government Interagency Working Group on Social Cost of Greenhouse Gases (2021), and we define $\chi = 8.887 \times 10^{-3}$ metric tons CO2 per gallon as the CO2 content of gasoline. The externality estimate is

$$\hat{\phi}_i = \left(\frac{m_i}{23} - \frac{m_i}{30}\right) \cdot \chi \cdot SCC. \tag{9}$$

[8]This calculation requires the assumption that depreciation is driven entirely by age, not mileage. Roughly consistent with that, Knittel and Sallee (2019) find that the effect of age on depreciation is about three times the effect of mileage.

3.1.2 Drinks Experiment

In the drinks experiment, we estimate biases related to imperfect information and self-control problems. We follow Allcott, Lockwood and Taubinsky (2019a) in constructing bias estimates from survey measures of nutrition knowledge and self-control. We define $\hat{\boldsymbol{b}}_i$ as a two-part vector containing participant i's nutrition knowledge and self-control variables from our survey. Let $\boldsymbol{b}^V$ be the nutrition knowledge and self-control of an unbiased consumer. Following Allcott, Lockwood and Taubinsky (2019a), we assume that unbiased consumers have nutrition knowledge of 0.92 (the average value for nutritionists and dieticians in their data) and self-control of 1 (i.e. they respond "not at all" when asked if they "drink soda pop or other sugar-sweetened beverages more often than I should"). Allcott, Lockwood and Taubinsky (2019a) show that under certain assumptions, the bias estimate is

$$\hat{\gamma}_i = \left(\boldsymbol{b}^V - \hat{\boldsymbol{b}}_i\right)\kappa, \tag{10}$$

where $\kappa > 0$ is an empirical constant.[9] Consistent with intuition, this equation assigns more bias to consumers with lower nutrition knowledge or lower self-control.

Again following Allcott, Lockwood and Taubinsky (2019a), we assume a relative externality from sugary drinks of $\hat{\phi} = 0.85$ cents per ounce.[10] We assume that this is constant across people and product pairs.

3.1.3 Bias and Externality Estimates

Figure 3 presents the distribution of bias estimates $\hat{\gamma}_{ij}$ for each experiment. This dispersion can reflect dispersion in (true) bias γ_{ij}, measurement error, or both. For the cars experiment, the average $\hat{\gamma}_{ij}$ is \$135 per year, and the standard deviation is \$735. The average gas cost savings from the higher-MPG car (given reported 2019 miles driven and gas prices) is about \$304 per year, so the average $\hat{\gamma}_{ij}$ implies inattention to about 44 percent of gas costs. This is more inattention than implied by other papers using market data or field experiments (Busse, Knittel and Zettelmeyer 2013; Allcott and Wozny 2014; Sallee, West and Fan 2016; Grigolon, Reynaert and Verboven 2018; Allcott and Knittel 2019). The average externality estimate $\hat{\phi}_{ij}$ is \$50 per year, as reported in a separate vertical line, and the standard deviation of $\hat{\phi}_{ij}$ is about \$22.

[9] Specifically, $\kappa = \hat{\boldsymbol{\lambda}} p^s / \hat{\zeta}^c$, where $\hat{\boldsymbol{\lambda}}$ is the association between bias proxies and sugary drink consumption, p^s is the price of sugary drinks, and $\hat{\zeta}^c > 0$ is the compensated elasticity of demand. We take $\hat{\boldsymbol{\lambda}} = [0.854, 0.825]$ from column 1 of Allcott, Lockwood, and Taubinsky's (2019b) Table V, and we take p^s = \$0.0363/ounce (\$5.23/12-pack) and $\hat{\zeta}^c = 1.39$ from their Table VI. Intuitively, bias $\hat{\gamma}_i$ (in dollar units) is larger if the bias proxies are more strongly associated with consumption ($\hat{\boldsymbol{\lambda}}$ is larger) or if a given consumption change corresponds to more dollars ($p/\hat{\zeta}^c$ is larger). See Allcott, Lockwood and Taubinsky (2019a) for discussion of required assumptions, including linearity and unconfoundedness.

[10] Their estimate of $\hat{\gamma}_i^e$ is derived from several earlier studies. Wang et al. (2012) use epidemiological simulation models to estimate that soda consumption increases health care costs by an average of approximately one cent per ounce. Yong et al. (2011) estimate that for people with employer-provided insurance, about 15 percent of health costs are borne by the individual, while 85 percent are covered by insurance. Similarly, Cawley and Meyerhoefer (2012) estimate that 88 percent of the total medical costs of obesity are borne by third parties, and obesity is one of the primary diseases thought to be caused by SSB consumption.

For the drinks experiment, the average $\hat{\gamma}_i$ is \$2.56 per 12-pack, and the standard deviation is about \$1.24. The relative externality $\hat{\phi}$ is a constant \$1.22 per 12-pack.

These results imply that nudges have beneficial average effects if they reduce demand for lower-MPG cars and sugary drinks, but the heterogeneity foreshadows the importance of reducing distortion variance.

3.2 Average Treatment Effect of Labels, Variance, and Covariance with Bias and Externalities

In this section, we estimate the average treatment effects of labels, the variance of treatment effects, and the covariance of treatment effects with bias and externalities.

3.2.1 Empirical Model

We define T_i as a treatment indicator, with $T_i = 1$ for a label treatment condition and $T_i = 0$ for control. (We pool all label conditions in some analyses, while in other analyses we compare an individual label condition against control.) Following the framework in Section 1, participant i's relative WTP for the more harmful good in product pair j elicited on choice occasion s depends on utility v_{ijs}, bias γ_{ij}, and nudge effect τ_{ij}.[11] We decompose the utility as $v_{ijs} = v_{ij} + \epsilon_{ijs}$, where v_{ij} is a stable person-specific component and ϵ_{ijs} is an idiosyncratic shock capturing taste variation, elicitation noise, or (in the cars experiment) the WTP difference between the Accord and Legacy. This implies that relative WTP is

$$w_{ijs} = v_{ij} + \gamma_{ij} + \tau_{ij} \cdot T_i \cdot \mathbf{1}[s = 2] + \epsilon_{ijs}. \tag{11}$$

We use the tilde ($\tilde{}$) to indicate differences between the baseline and endline choice occasions. For example, $\tilde{w}_{ij} \coloneqq w_{ij2} - w_{ij1}$. Differencing equation (11) gives

$$\tilde{w}_{ij} = \tau_{ij} \cdot T_i + \tilde{\epsilon}_{ij}. \tag{12}$$

We make the following assumptions, which deliver a standard mixed effects model.

Assumption 1. The measurement errors in bias and externality, $\hat{\gamma} - \gamma$ and $\hat{\phi} - \phi$, are mean-zero, normally distributed, and independent of γ, ϕ, and τ.

Assumption 2. The treatment is randomly assigned: $T \perp \tilde{\epsilon}$.

Assumption 3. The idiosyncratic WTP change is orthogonal to the treatment effect : $\tilde{\epsilon} \perp \tau$.

Assumption 4. τ, γ, ϕ, and $\tilde{\epsilon}$ are normally distributed.

Perhaps the strongest assumption is the part of Assumption 1 that requires measurement error $\hat{\gamma} - \gamma$ to be independent of treatment effects τ. For example, this implies that the treatment effects

[11] In the cars experiment, recall that we randomized whether the Accord or Legacy was valued at baseline instead of endline. Accord-Fusion at baseline followed by Legacy-Fusion at endline is defined as a separate product-pair from Legacy-Fusion followed by Accord-Fusion.

of nutrition facts labels don't covary with misperceptions of calorie or sugar content that are not predicted by our nutrition knowledge and self-control bias proxies. In the cars experiment, we have an error-free measure of failure to optimize within the experiment, but Assumption 1 would be required to consider potential bias in actual car purchase decisions.

An alternative strategy in Appendix D.2 delivers $Cov[\gamma, \tau]$ in the drinks experiment without assuming normality; the estimate is very similar.

Assumptions 1 and 4 together imply that $\hat{\gamma}$ and $\hat{\phi}$ are normally distributed. Since τ is also normal, τ_{ij} can be written as

$$\tau_{ij} = \eta_{ij} + \alpha_1 \hat{\gamma}_{ij} + \alpha_2 \hat{\phi}_{ij}, \tag{13}$$

where η is normal and independent of $\left(\hat{\gamma}, \hat{\phi}\right)$, $\alpha_1 = Cov[\hat{\gamma}, \tau] / Var[\hat{\gamma}]$, and $\alpha_2 = Cov\left[\hat{\phi}, \tau\right] / Var\left[\hat{\phi}\right]$. By Assumption 1, $Cov[\hat{\gamma}, \tau] = Cov[\gamma, \tau]$ and $Cov\left[\hat{\phi}, \tau\right] = Cov[\phi, \tau]$, and thus

$$Cov[\gamma, \tau] = \alpha_1 Var[\hat{\gamma}] + \alpha_2 Cov\left[\hat{\gamma}, \hat{\phi}\right] \tag{14}$$

$$Cov[\phi, \tau] = \alpha_2 Var\left[\hat{\phi}\right] + \alpha_1 Cov\left[\hat{\gamma}, \hat{\phi}\right]. \tag{15}$$

Combining equations (12) and (13) gives

$$\tilde{w}_{ij} = \eta_{ij} T_i + \alpha_1 \hat{\gamma}_{ij} T_i + \alpha_2 \hat{\phi}_{ij} T_i + \tilde{\epsilon}_{ij}. \tag{16}$$

A linear projection of $\tilde{\epsilon}_{ij}$ on $\hat{\gamma}_{ij}$ and $\hat{\phi}_{ij}$ gives

$$\tilde{\epsilon}_{ij} = \beta_1 \hat{\gamma}_{ij} + \beta_2 \hat{\phi}_{ij} + \beta_{0i} + \nu_{ij}, \tag{17}$$

where β_{0i} is an individual-level constant and the residual ν_{ij} is mean-zero.[12] Our assumptions imply that $\beta_{0i} + \nu_{ij}$ is normally distributed, and we additionally assume that β_{0i} and ν_{ij} are each normally distributed. Combining equations (16) and (17) gives a mixed effects model:

$$\tilde{w}_{ij} = \eta_{ij} T_i + \alpha_1 \hat{\gamma}_{ij} T_i + \alpha_2 \hat{\phi}_{ij} T_i + \beta_1 \hat{\gamma}_{ij} + \beta_2 \hat{\phi}_{ij} + \beta_{0i} + \nu_{ij}. \tag{18}$$

Using Assumptions 1 and 2, we estimate $\mathbb{E}[\tau]$ using equation (12) with fixed τ in ordinary least squares, i.e. by simply regressing $\tilde{w}_{ij}$ on T_i and a constant. Using orthogonality and normality from Assumptions 2 and 3, we estimate $Var[\tau]$ using equation (12) in a standard mixed effects maximum likelihood estimator. Using orthogonality and normality from Assumptions 1–4, we also estimate equation (18) via mixed effects. Then, using the estimated regression coefficients $\{\alpha_1, \alpha_2\}$ and the variances and covariance $\left\{Var[\hat{\gamma}], Var\left[\hat{\phi}\right], Cov\left[\hat{\gamma}, \hat{\phi}\right]\right\}$ computed directly from the sample distribution of bias and externality estimates, we then recover the covariances $\{Cov[\gamma, \tau], Cov[\phi, \tau]\}$

[12] The coefficients are $\beta_1 = Cov[\hat{\gamma}, \tilde{\epsilon}] / Var[\hat{\gamma}]$ and $\beta_2 = Cov\left[\hat{\phi}, \tilde{\epsilon}\right] / Var\left[\hat{\phi}\right]$.

using equations (14) and (15).

Because equations (12)–(18) also hold conditional on controls, we can improve precision by adding controls. In all regressions, we include indicators for product pairs j and MPL order.

3.2.2 Descriptive Figures and Intuition for Identification

Before presenting formal regression results, we present figures that illustrate the identification. Figure 4 presents the treatment and control group distributions of WTP changes $\tilde{w}_{ij}$ across people and product pairs in each experiment, pooling treatment group data across all labels. In both experiments, the treatment distributions are shifted to the left, implying that the labels reduced WTP for lower-MPG cars (relative to higher-MPG cars) and for sugary drinks (relative to sugar-free drinks). The treatment distributions are also more dispersed, with less weight on \$0 change and more weight throughout the left tails. In both experiments, we should thus expect $\mathbb{E}[\tau] < 0$ and a significant $Var[\tau]$, since the effect of the label on dispersion of $\tilde{w}$ is what identifies $Var[\tau]$.

Figure 5 presents average treatment effect (ATE) estimates using the OLS version of equation (12), for each label in each experiment. The left-most bars for each label are the average treatment effects $\mathbb{E}[\tau]$. The remaining bars present conditional ATEs within the sample of participant-by-product pair observations with below-median or above-median bias or externality estimates. Holding the ATE and $Var[\tau]$ constant, a more negative effect for above-median distortions $\hat{\delta}$ would imply a more negative $Cov[\delta, \tau]$, and thus a reduction in the variance of distortions.

For the cars experiment, the ATEs and conditional ATEs are statistically indistinguishable across labels. For the personalized cost label, the cost differences presented are larger for heavy drivers and smaller for light drivers, which could have generated larger effect sizes for heavy drivers, analogous to Davis and Metcalf (2016). Our point estimates are consistent with this, but the difference is highly statistically insignificant.

For the drinks experiment, there is some suggestive evidence of tradeoffs. The graphic warning label has the largest ATE, but it has smaller effects for more biased consumers, implying a positive $Cov[\hat{\gamma}, \tau]$. The nutrition facts label has smaller ATE but larger point estimates for more biased consumers. These differences are not statistically significant, but they suggest that the nutrition facts label might generate larger total surplus gains despite having smaller average effects. When pooling across the three sugary drink labels in the right-most set of results, the point estimates suggest slightly smaller effects for more biased consumers, implying slightly positive $Cov[\hat{\gamma}, \tau]$. Recall that for sugary drinks, we assume that the externality ϕ is constant across consumers.

Appendix Figure A8 analyzes treatment effect heterogeneity separately for the two types of bias proxies. Although not statistically different, the point estimates suggest that the labels have larger effects for consumers with below-median nutrition knowledge, which pushes toward decreasing the variance of distortions. However, the labels have significantly smaller effects for consumers with below-median self-control, which pushes towards increasing the variance of distortions. This is especially true for the graphic warning labels: those effects are almost three times smaller for people who say they "definitely" or "mostly" drink sugary drinks "more often that I should" than

they are for people who say that they only "somewhat" or "not at all" drink more than they should. In other words, graphic warning labels have the smallest effects on the people whose survey responses suggest they want the most help.

3.2.3 Parameter Estimates

Table 1 presents results with data pooled across all labels. Columns 1 and 2 present fixed coefficient (OLS) versions of equations (12) and (16), respectively, while columns 3 and 4 present the full random coefficient (mixed effects) models. The coefficients are very similar between OLS and mixed effects.

From column 1, the average treatment effect $\mathbb{E}[\tau]$ for fuel economy labels is a reduction in relative WTP for the lower-MPG car of \$59 per vehicle-year. Using the demand slope reported in Section 4.2, a homogeneous treatment effect of that magnitude would reduce the lower-MPG car's market share by about $\mathbb{E}[\tau] D_p' \approx 4$ percentage points. This is within the confidence intervals of the (statistically insignificant) effects of fuel cost information provision on actual vehicle purchases from the experiment in Allcott and Knittel (2019).

The ATE for sugary drink labels is a reduction in relative WTP for the sugary drink of \$0.43 per 12-pack. Using the demand slope, a homogeneous treatment effect of that magnitude would reduce the sugary drink's market share by about 6 percentage points. This is somewhat smaller than the 10 and 17 percentage point effects of stop sign and graphic warning labels on sugary drink demand in Grummon et al. (2019a) and Hall et al. (2022), respectively.

Both fuel economy and health labels have beneficial average effects: the average effects offset the suboptimally high purchases of low-MPG vehicles and sugary drinks caused by bias and externalities. This offset is only partial: labels offset $59/(135+50) \approx 32\%$ of the bias + externality in the cars experiment, and labels offset only $0.43/(2.56+1.22) \approx 11\%$ in the drinks experiment. This is consistent with the view that in markets with material externality problems, nudges don't generate enough behavior change to eliminate the need for taxation (e.g., Loewenstein and Chater 2017; Thaler and Sunstein 2021).

From column 3, the variance of fuel economy label effects is $Var[\tau] \approx$ 30,428 (\$/vehicle-year)^2, implying a standard deviation of \$174 per vehicle-year, and thus a coefficient of variation ($\sqrt{Var[\tau]}/\mathbb{E}[\tau]$) of about 3.0. The variance of sugary drink label effects is 0.74 (\$/12-pack)^2, implying a standard deviation of \$0.86 and a coefficient of variation of about 2.0.

Column 4 presents the primary estimate of $\{\alpha_1, \alpha_2\}$, the coefficients on bias or externality interacted with T_i, in mixed effects. For the cars experiment, $\alpha_1 \approx -0.01$ and $\alpha_2 \approx -0.007$, both of which are statistically indistinguishable from zero, and $Cov\left[\hat{\gamma}, \hat{\phi}\right] \approx$ 2,384. From Section 3.1, the standard deviations of $\hat{\gamma}$ and $\hat{\phi}$ are \$735 and \$22 per year. Thus, using equations (14) and (15), the covariance point estimates are $Cov[\gamma, \tau] \approx$ -7,744 and $Cov[\phi, \tau] \approx$ -37 (\$/vehicle)^2.

For the drinks experiment, $\alpha_1 \approx 0.08$, which is statistically significantly different from zero. From Section 3.1, the standard deviation of $\hat{\gamma}$ is \$1.24 per 12-pack. Thus, the covariance point estimate is $Cov[\gamma, \tau] \approx 0.08 \times 1.24^2 \approx 0.13$ (\$/12-pack)^2.

Appendix D.1 presents versions of Table 1 specific to each of the seven labels across the two experiments. One key result that differs across labels is that the graphic warning label has significantly more positive $Cov[\gamma, \tau]$ than the nutrition facts label, consistent with the suggestive graphical results in Figure 5.

3.3 Psychic Benefits and Costs

For the cars experiment, the labels are not very evocative, so we assume zero psychic benefit ($\Delta I = 0$). For the drinks experiment, we use the MPL questions eliciting hypothetical WTP to receive or avoid receiving labels on a 12-pack of sugary drinks. We assume that those who buy sugar-free drinks experience no psychic benefits or costs ($\iota_0 = 0$), so aggregate psychic benefits with the labels are $I(\sigma, p) = \sigma \bar{\iota}_1 q^*(\sigma = 1)$, where $\bar{\iota}_1$ is the average psychic benefit conditional on buying the sugary drink, and $q^*(\sigma = 1)$ is the equilibrium quantity. ι_1 could be negative if the label is disgusting or otherwise aversive, or it could be positive if consumers value the hard information or reminders about health even after buying the drinks.

Figure 6 presents average WTP by label. The average person in our sample wants to receive the nutrition label and reported being willing to pay \$2 for it on a 12-pack. Average WTP is also positive for the stop sign label. The average person in our sample reported being willing to pay about \$1 to avoid receiving a 12-pack with the graphic warning labels. Pooling across all three labels, average WTP is almost exactly zero.

Appendix Figure A9 presents the distribution of reasons why people did or did not want to receive each label, as elicited from multiple choice questions. People who wanted to receive the nutrition and stop sign labels mostly reported that this was because "the information on the label is useful," while the minority of people who wanted to receive the graphic warning label mostly reported that this was because "it would remind me to drink less." The modal person who wanted to avoid receiving the labels reported that this was because "the label makes me feel bad," although a large minority of people reported that "I don't need the government to tell me what to drink." Another large minority of people who wanted to avoid the graphic warning labels reported that this was because "the label is gross."

4 Welfare Analysis

4.1 Implementation of Proposition 1.3

In our empirical applications, we consider a binary nudge intensity, $\sigma \in \{0, 1\}$, corresponding to the question of whether or not to have labels. To easily apply Proposition 1 to this case, we assume that the moments are locally linear in σ.[13] We define a difference operator $\Delta X(\sigma) \coloneqq X(\sigma = 1) - X(\sigma = 0)$ for any function $X(\sigma)$.

[13] Without this assumption, we would need to integrate the first-order condition in Proposition 1 across all values of $\sigma \in [0, 1]$.

As noted in Section 3, we increase precision and generalizability by assuming that the population parameters approximate the marginal parameters.

We explicitly add an externality ϕ to the bias γ, and we define $\delta \coloneqq \gamma + \phi$ as their sum. Because covariances are bilinear, $Cov[\delta, \tau] = Cov[\gamma, \tau] + Cov[\phi, \tau]$. As described in Section 1.1, Proposition 1 holds after replacing γ with δ. We also assume zero psychic benefits/costs ($I = 0$) for the quantitative analyses.

When we make the assumptions above, expand the second moment, and rearrange, Proposition 1 implies that the total surplus changes from the nudge at $t = 0$ and at optimal tax $t^* = \mathbb{E}[\delta + \sigma\tau] - \mu$, respectively, are:

$$\Delta W(t = 0) \approx \frac{1}{2}\left\{\rho\left(\left(\mathbb{E}[\delta] - \mu + \mathbb{E}[\tau]\right)^2 - \left(\mathbb{E}[\delta] - \mu\right)^2\right) + 2Cov[\delta, \tau] + Var[\tau]\right\} D'_p \tag{19}$$

$$\Delta W(t = t^*) \approx \frac{1}{2}\left\{2Cov[\delta, \tau] + Var[\tau]\right\} D'_p. \tag{20}$$

We take $\mathbb{E}[\delta]$, $\mathbb{E}[\tau]$, $Var[\tau]$, and $Cov[\delta, \tau]$ from Section 3. Below, we discuss the remaining parameters and then present total surplus effects.

4.2 Demand Slope

We pool across all product pairs to construct an average demand function for the more harmful good, as illustrated in Appendix Figure A7. The average demand slopes for the fuel economy and drinks experiments are -0.00060 market share per \$/vehicle-year and -0.14 market share per \$/12-pack, respectively.

4.3 Pass-Through and Markup Assumptions

When there are two or more goods, it is the difference in markups and pass-through rates that form the sufficient statistics for welfare analysis; see Appendix A.4. The pass-through rate ρ is the pass-through of a tax on the more harmful good to the price of the more harmful good relative to the less harmful good. The markup μ is the relative markup on the more versus less harmful good.

4.3.1 Cars Experiment

There are no papers that specifically study how taxes on 23-MPG sedans are passed through to those sedans vs. 30-MPG sedans. As a proxy, we collect estimates from three papers studying pass-through rates for hybrid vehicle subsidies and manufacturers' customer rebates, which use the prices of other vehicles as controls.[14] The average pass-through rate is approximately 0.8, so we assume $\rho \approx 0.8$ as a rough benchmark.

[14]The papers are Busse, Silva-Risso and Zettelmeyer (2006), Sallee (2011), and Gulati, McAusland and Sallee (2017).

Since prices and cost structures are similar for our 23-MPG versus 30-MPG sedans, we assume $\mu = 0$.

4.3.2 Drinks Experiment

For sugary drinks, we collect estimates from seven papers that study city-level sugary drink taxes in the U.S.[15] The average pass-through rate to taxed beverages is 0.85, and the average pass-through rate to untaxed beverages is 0.05, giving $\rho \approx 0.85 - 0.05 \approx 0.80$.

Three of these seven papers report prices for both sugary drinks and water.[16] The prices are relatively similar. While there is no public information on the marginal costs of 12-packs of sugary drinks versus sparkling water, they have very similar cost structures. Thus, we assume similar markups on the two goods, implying a relative markup of $\mu = 0$.

4.4 Results

Table 2 summarizes the sufficient statistics and resulting total surplus effects obtained from equations (19) and (20). We initially pool across labels in each experiment to simplify and increase precision.

To demonstrate the computation, consider the total surplus effect of fuel economy labels with no tax. Using $\mu = 0$, $\mathbb{E}[\delta] = \mathbb{E}[\gamma] + \mathbb{E}[\phi] = 135 + 50$, and $Cov[\delta, \tau] = Cov[\gamma, \tau] + Cov[\phi, \tau] =$ -7,744 + −37, equation (19) is

$$\begin{aligned}
\Delta W(t=0) \approx & \frac{1}{2}\left\{\rho\left(\left(\mathbb{E}[\delta] + \mathbb{E}[\tau]\right)^2 - \left(\mathbb{E}[\delta]\right)^2\right) + 2Cov[\delta, \tau] + Var[\tau]\right\} D_p' && (21)\\
\approx & \frac{1}{2}\left\{0.80\left(\ (184 + -59)^2\ \ - (184)^2\right) + 2(-7,781) + 30,428\right\}(-0.00060) && (22)\\
\approx & -0.07\ \ \$/\text{vehicle-year}.
\end{aligned}$$

Thus, the point estimates suggest that fuel economy labels decrease total surplus in our experiment.

Three special cases reported near the bottom of Table 2 illustrate the importance of the three terms inside brackets, which correspond to the three economic forces described in Section 1.3. First, labels with homogeneous effects ($Cov[\delta, \tau] = Var[\tau] = 0$) and the same estimated $\mathbb{E}[\tau]$ would increase total surplus by \$4.36 per vehicle-year. Second, labels with zero average effect ($\mathbb{E}[\tau] = 0$) and the same estimated $Cov[\delta, \tau]$ and $Var[\tau]$ would decrease total surplus by \$4.43 per vehicle-year. Third, "pure noise" labels with zero average effect and $Cov[\delta, \tau] = 0$ and the same estimated $Var[\tau]$ would decrease total surplus by \$9.07 per vehicle-year. The first case shows the importance of average behavior change, which drives the overall total surplus gain. The latter two cases show that the labels add substantial noise to consumer choice in our experiment, and this misallocation is quantitatively important for welfare.

[15] The papers are Falbe et al. (2015), Silver et al. (2017), Cawley et al. (2018), Cawley et al. (2020), Bleich et al. (2021), Seiler, Tuchman and Yao (2021), Petimar et al. (2022)

[16] The three papers are Falbe et al. (2015), Cawley et al. (2020), and Seiler, Tuchman and Yao (2021).

The final row of Table 2 shows that fuel economy labels reduce total surplus when they are paired with the optimal tax. Intuitively, since the tax allows the government to optimally control aggregate quantity, the labels' average effect is irrelevant, and the increase in distortion variance reduces total surplus. In all cases, the estimated total surplus effects are small relative to lease prices of \$5,600 per vehicle-year (Brozic 2022). This is driven by the modest demand slope D'_p and small average treatment effects.

In the drinks experiment, the point estimates of the variance and covariance are much smaller compared to the bias and externality. With no tax, health labels increase total surplus in our model, again primarily because the labels have beneficial average effects. At the optimal tax, the labels reduce total surplus in our model because they increase the distortion variance, due to their high variance and positive covariance with bias.

Figures 7 and 8 illustrate how the total surplus effects of labels and taxes depend on different parameter values. Assuming locally linear demand, the total surplus gain from the optimal tax t^* is $-\frac{1}{2}t^{*2}D'_p$. In both experiments, the labels reduce the net distortion, so taxes and labels are substitutes: total surplus gains from optimal taxes are smaller with the labels, and total surplus gains from labels are smaller with optimal taxes. Total surplus gains from optimal taxes depend only on D'_p and the average marginal net distortion $\mathbb{E}[\delta + \sigma\tau] - \mu$, so those lines are flat in Figure 7 with respect to $Var[\tau]$ and $Cov[\delta, \tau]$.

The left panels in Figure 7 illustrate one basic insight from Proposition 1 by plotting ΔW as a function of $Var[\tau]$, holding constant the other parameters in Table 2. The total surplus gains from labels are declining in $Var[\tau]$ because the variance of distortions is lower at lower values of $Var[\tau]$, all else equal. In both experiments, total surplus gains differ significantly for plausible values of $Var[\tau]$. In the cars experiment, the negative $Cov[\delta, \tau]$ implies that fuel economy labels might even be preferred to optimal taxes at small values of $Var[\tau]$. By contrast, at twice the estimated $Var[\tau]$, fuel economy labels generate substantial misallocation, reducing total surplus by about the amount that the optimal tax would increase it.

In the drinks experiment, optimal taxes are preferred to labels at most plausible values because the average net distortion is large relative to the average treatment effect of the labels. In the drinks experiment at $t = 0$, labels increase total surplus at small values of $Var[\tau]$ because the decrease in average demand reduces the average net distortion, but they decrease total surplus at larger values of $Var[\tau]$.

The right panels in Figure 7 plot ΔW as a function of $Cov[\delta, \tau]$, holding constant the other parameters. The total surplus gains from labels are declining in $Cov[\delta, \tau]$, because a more negative $Cov[\delta, \tau]$ implies that a larger share of the assumed variance of treatment effects is well-targeted. In both experiments, plausible differences in $Cov[\delta, \tau]$ again substantially affect total surplus. In the cars experiment, if $Cov[\delta, \tau]$ were twice as large as our point estimates, fuel economy labels would be preferred to optimal taxes. If $Cov[\delta, \tau]$ had the same magnitude but opposite sign, the fuel economy labels would significantly reduce total surplus.

The left panels in Figure 8 show how total surplus effects depend on pass-through assumptions.

At $\rho = 0$, implying perfectly inelastic supply, taxes do not affect equilibrium quantity or total surplus. The optimal tax increases total surplus as ρ increases, especially without labels, again because optimal taxes and labels are substitutes at our parameter values. At the optimal tax, pass-through does not affect the total surplus effect of labels because the government uses the tax to optimally set aggregate quantity. At $t = 0$, pass-through does matter because the labels decrease average demand, offsetting the distortion from bias and externalities. Our three auto market pass-through estimates (from footnote 14) range from 0.57 to 1, while different combinations of estimates from the seven soda tax papers (from footnote 15) could imply relative pass-through from $\rho = 0.16$ to $\rho = 1.37$. In both experiments when $t = 0$, using estimates from the lower versus higher ends of those ranges would significantly change the magnitude of total surplus effects and could even change the sign.

The right panels in Figure 8 show how the markup affects total surplus. To ease interpretation, we divide the markup difference μ by benchmark estimates of average prices of each good: \$5,600 to lease a car for one year (Brozic 2022) and \$4 for a 12-pack of drinks. The welfare effects of optimal taxes are quadratic in μ because μ directly determines the net distortion, and the minimum total surplus effect is zero at $t^* = 0$. At the optimal tax, the total surplus effects of labels do not depend on the markup, because the tax rate can be set to offset any markup difference. At $t = 0$, the total surplus effect of labels is decreasing in μ, as the average demand decrease caused by the labels is more beneficial with an additional pre-existing distortion that decreases the price of the more harmful good.

Appendix Table A13 presents label-specific point estimates. The parameters and total surplus effects are mostly statistically indistinguishable across labels. The point estimates suggest that although graphic warning labels on sugary drinks have larger ATEs than the nutrition facts labels, the increased distortion variance (larger $Var[\tau]$ and significantly larger $Cov[\gamma, \tau]$) implies that the graphic labels generate lower total surplus gains. Adding the large psychic benefits and costs quantified in Section 3.3 would imply large total surplus gains from nutrition facts labels and large total surplus losses from graphic warning labels.

5 Conclusion

Simplified information provision, warning labels, and other "nudges" have become increasingly popular over the past 15 years. Much of the public discussion and empirical work focuses on whether these instruments have beneficial average effects, and the original ideas of asymmetric paternalism and libertarian paternalism have not been formalized in market settings. Our theoretical framework fleshes out these ideas and clarifies how average effects, asymmetric or libertarian paternalism, deliberative competence, and other concepts matter for welfare. Our randomized experiments demonstrate both the challenges and opportunities in empirically evaluating nudges in a theoretically grounded framework. The results highlight the importance of measuring both aversiveness and targeting, illustrating how average effects can be an incomplete and possibly misleading proxy

for welfare.

References

Allcott, Hunt, "Social Norms and Energy Conservation," *Journal of Public Economics*, 2011, *95* (9-10), 1082–1095.

_ , "Site Selection Bias in Program Evaluation," *Quarterly Journal of Economics*, 2015, *130* (3), 1117–1165.

_ **and Christopher Knittel**, "Are Consumers Poorly Informed about Fuel Economy? Evidence from Two Experiments," *American Economic Journal: Economic Policy*, 2019, *11* (1), 1–37.

_ **and Dmitry Taubinsky**, "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market," *American Economic Review*, 2015, *105* (8), 2501–38.

_ **and Judd B Kessler**, "The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons," *American Economic Journal: Applied Economics*, 2019, *11* (1), 236–76.

_ **and Nathan Wozny**, "Gasoline Prices, Fuel Economy, and the Energy Paradox," *Review of Economics and Statistics*, 2014, *96* (5), 779–795.

_ **and Richard L. Sweeney**, "The Role of Sales Agents in Information Disclosure: Evidence from a Field Experiment," *Management Science*, 2017, *63* (1), 21–39.

_ **and Todd Rogers**, "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation," *American Economic Review*, 2014, *104* (10), 3003–37.

_ **, Benjamin B Lockwood, and Dmitry Taubinsky**, "Regressive Sin Taxes, With an Application to the Optimal Soda Tax," *Quarterly Journal of Economics*, 2019, *134* (3), 1557–1626.

_ **, Benjamin B. Lockwood, and Dmitry Taubinsky**, "Should We Tax Sugar-Sweetened Beverages? An Overview of Theory and Evidence," *Journal of Economic Perspectives*, 2019, *33* (3), 202–27.

Allende, Claudia, Francisco Gallego, and Christopher Neilson, "Approximating The Equilibrium Effects of Informed School Choice," Working Paper 2019-16, Princeton University Economics Department. 2019.

Ambuehl, Sandro, B. Douglas Bernheim, and Annamaria Lusardi, "Evaluating Deliberative Competence: A Simple Method with an Application to Financial Choice," *American Economic Review*, 2022, *112* (11), 3584–3626.

Araya, Sebastian, Andres Elberg, Carlos Noton, and Daniel Schwartz, "Identifying Food Labeling Effects on Consumer Behavior," *Marketing Science*, 2022, *41* (5).

Barahona, Nano, Cristobal Otero, and Sebastian Otero, "Equilibrium Effects of Food Labeling Policies," *Econometrica*, 2023, *91* (3), 839–868.

_ , _ , _ , **and Joshua Kim**, "On the Design of Food Labeling Policies," *Available at SSRN 4079728*, 2022.

Benartzi, Shlomo, John Beshears, Katherine L. Milkman, Cass R. Sunstein, Richard H. Thaler, Maya Shankar, Will Tucker-Ray, William J. Congdon, and Steven Galing, "Should Governments Invest More in Nudging?," *Psychological Science*, 2017, *28* (8), 1041–1055.

Bernheim, B Douglas and Dmitry Taubinsky, "Behavioral Public Economics," in "Handbook of Behavioral Economics: Applications and Foundations," Vol. 1, Elsevier, 2018, pp. 381–516.

_ **, Andrey Fradkin, and Igor Popov**, "The Welfare Economics of Default Options in 401(k) Plans," *American Economic Review*, 2015, *105* (9), 2798–2837.

Beshears, John, James J Choi, David Laibson, Brigitte C Madrian, and Katherine L Milkman, "The Effect of Providing Peer Information on Retirement Savings Decisions," *Journal of Finance*, 2015, *70* (3), 1161–1201.

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

Bleich, Sara N., Caroline G. Dunn, Mark J. Soto, Jiali Yan, Laura A. Gibson, Hannah G. Lawman, Nandita Mitra, Caitlin M. Lowery, Ana Peterhans, Sophia V. Hua, and Christina A. Roberto, "Association of a Sweetened Beverage Tax With Purchases of Beverages and High-Sugar Foods at Independent Stores in Philadelphia," 2021, *4* (6), e2113527–e2113527.

Bollinger, Bryan, Phillip Leslie, and Alan Sorensen, "Calorie Posting in Chain Restaurants," *American Economic Journal: Economic Policy*, 2011, *3* (1), 91–128.

Bond, Samule D., Kurt A. Carlson, and Ralph L. Keeney, "Generating Objectives: Can Decision Makers Articulate What They Want?," *Management Science*, 2008, *54* (1), 56–70.

Brewer, Noel T, Marissa G Hall, Seth M Noar, Humberto Parada, Al Stein-Seroussi, Laura E Bach, Sean Hanley, and Kurt M Ribisl, "Effect of Pictorial Cigarette Pack Warnings on Changes in Smoking Behavior: A Randomized Clinical Trial," *JAMA Internal Medicine*, 2016, *176* (7), 905–912.

Brozic, Jennifer, "How Much Does it Cost to Lease a Car?," *Credit Karma*, 2022. Available at `https://www.creditkarma.com/auto/i/cost-to-lease-a-car`.

Busse, Meghan, Jorge Silva-Risso, and Florian Zettelmeyer, "$1,000 Cash Back: The Pass-Through of Auto Manufacturer Promotions," *American Economic Review*, 2006, *96* (4), 1253–1270.

Busse, Meghan R., Christopher R. Knittel, and Florian Zettelmeyer, "Are Consumers Myopic? Evidence from New and Used Car Purchases," *American Economic Review*, 2013, *103* (1), 220–56.

Butera, Luigi, Robert Metcalfe, William Morrison, and Dmitry Taubinsky, "Measuring the Welfare Effects of Shame and Pride," *American Economic Review*, 2022, *112* (1), 122–168.

Camerer, Colin, Samuel Issacharoff, George Loewenstein, Ted O'Donoghue, and Matthew Rabin, "Regulation for Conservatives: Behavioral Economics and the Case for 'Asymmetric Paternalism'," *University of Pennsylvania Law Review*, 2003, *151* (3), 1211–1254.

Cantrell, Jennifer, Donna M Vallone, James F Thrasher, Rebekah H Nagler, Shari P Feirman, Larry R Muenz, David Y He, and Kasisomayajula Viswanath, "Impact of Tobacco-Related Health Warning Labels Across Socioeconomic, Race and Ethnic Groups: Results from a Randomized Web-based Experiment," *PLoS One*, 2013, *8* (1), e52206.

Carroll, Gabriel D, James J Choi, David Laibson, Brigitte C Madrian, and Andrew Metrick, "Optimal Defaults and Active Decisions," *Quarterly Journal of Economics*, 2009, *124* (4), 1639–1674.

Cawley, John and Chad Meyerhoefer, "The Medical Care Costs of Obesity: An Instrumental Variables Approach," *Journal of Health Economics*, 2012, *31* (1), 219–230.

_ , Chelsea Crain, David Frisvold, and David Jones, "The Pass-Through of the Largest Tax on Sugar-Sweetened Beverages: The Case of Boulder, Colorado," *NBER Working Paper*, 2018. NBER WP No. 25050.

_ , David Frisvold, Anna Hill, and David Jones, "The Impact of the Philadelphia Beverage Tax on Prices and Product Availability," *Journal of Policy Analysis and Management*, 2020, *39* (3), 605–628.

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

_ , **John N. Friedman, Soren Leth-Petersen, Torben Heien Nielsen, and Tore Olsen**, "Active vs. Passive Decisions and Crowd-Out in Retirement Savings Accounts: Evidence from Denmark," 2014, *129* (3), 1141–1219.

Choukhmane, Taha and Christopher Palmer, "How Do Consumers Finance Increased Retirement Savings?," *working paper*, 2023.

Damgaard, Mette Trier and Christina Gravert, "Now or never! The effect of deadlines on charitable giving: Evidence from two natural field experiments," *Journal of Behavioral and Experimental Economics*, 2016, *66*.

_ **and** _ , "The hidden costs of nudging: Experimental evidence from reminders in fundraising," *Journal of Public Economics*, 2018, *157*, 15–26.

Davis, Lucas W and Gilbert E Metcalf, "Does Better Information Lead to Better Choices? Evidence from Energy-efficiency Labels," *Journal of the Association of Environmental and Resource Economists*, 2016, *3* (3), 589–625.

Delipalla, Sofia and Michael Keen, "The Comparison Between Ad Valorem and Specific Taxation Under Imperfect Competition," *Journal of Public Economics*, 1992, *49* (3), 351–367.

DellaVigna, Stefano and Elizabeth Linos, "RCTs to Scale: Comprehensive Evidence from Two Nudge Units," *Econometrica*, 2022, *90*, 81–116.

Delmas, Magali A., Miriam Fischlein, and Omar I. Asensio, "Information Strategies and Energy Conservation Behavior: A Meta-analysis of Experimental Studies From 1975 to 2012," *Energy Policy*, 2013, *61*, 729–739.

Donnelly, Grant E, Laura Y Zatz, Dan Svirsky, and Leslie K John, "The Effect of Graphic Warnings on Sugary-Drink Purchasing," *Psychological Science*, 2018, *29* (8), 1321–1333.

Exley, Christine L and Judd B Kessler, "Motivated Errors," *NBER Working Paper*, 2019. NBER WP No. 26595.

Falbe, Jennifer, Nadia Rojas, Anna H. Grummon, and Kristine A. Madsen, "Higher Retail Prices of Sugar-Sweetened Beverages 3 Months After Implementation of an Excise Tax in Berkeley, California," *American Journal of Public Health*, 2015, *105* (11), 2194–2201. PMID: 26444622.

Farhi, Emmanuel and Xavier Gabaix, "Optimal Taxation with Behavioral Agents," *American Economic Review*, 2020, *110* (1), 298–336.

Finkelstein, Amy and Matthew J Notowidigdo, "Take-Up and Targeting: Experimental Evidence from SNAP," *Quarterly Journal of Economics*, 2019, *134* (3), 1505–1556.

Gerber, Alan S and Todd Rogers, "Descriptive Social Norms and Motivation to Vote: Everybody's Voting and So Should You," *Journal of Politics*, 2009, *71* (1), 178–191.

Gillingham, Kenneth, Sebastien Houde, and Arthur van Benthem, "Consumer Myopia in Vehicle Purchases: Evidence from a Natural Experiment," *NBER Working Paper*, 2019. NBER WP No. 25845.

Gine, Xavier, Dean Karlan, and Jonathan Zinman, "Put Your Money Where Your Butt Is: A Commitment Contract for Smoking Cessation," *American Economic Journal: Applied Economics*, 2010, *2* (4), 213–35.

Goldin, Jacob and Daniel Reck, "Revealed-Preference Analysis with Framing Effects," *Journal of Political Economy*, 2020, *128* (7), 2759–2795.

Grigolon, Laura, Mathias Reynaert, and Frank Verboven, "Consumer Valuation of Fuel Costs and Tax Policy: Evidence from the European Car Market," *American Economic Journal: Economic Policy*, 2018, *10* (3), 193–225.

Grummon, Anna H and Marissa G Hall, "Sugary Drink Warnings: A Meta-Analysis of Experimental Studies," *PLoS Medicine*, 2020, *17* (5), e1003120.

Grummon, Anna H., Lindsey S. Taillie, Shelley D. Golden, Marissa G. Hall, Leah M. Ranney, and Noel T. Brewer, "Sugar-Sweetened Beverage Health Warnings and Purchases: A Randomized Controlled Trial," *American Journal of Preventive Medicine*, 2019, *57* (5), 601–610.

Grummon, Anna H, Marissa G Hall, Lindsey Smith Taillie, and Noel T Brewer, "How Should Sugar-sweetened Beverage Health Warnings Be Designed? A Randomized Experiment," *Preventive Medicine*, 2019, *121*, 158–166.

Gulati, Sumeet, Carol McAusland, and James M. Sallee, "Tax Incidence with Endogenous Quality and Costly Bargaining: Theory and Evidence From Hybrid Vehicle Subsidies," *Journal of Public Economics*, 2017, *155*, 93–107.

Hall, Marissa G., Anna H. Grummon, Isabella C. A. Higgins, Allison J. Lazard, Carmen E. Prestemon, Mirian I. Avendano-Galdamez, and Lindsey Smith Taillie, "The Impact of Pictorial Health Warnings on Purchases of Sugary Drinks for Children: A Randomized Controlled Trial," *PLOS Medicine*, 2022, *19* (2), 1–18.

Hammond, David, James Thrasher, Jessica L Reid, Pete Driezen, Christian Boudreau, and Edna Arillo Santillán, "Perceived Effectiveness of Pictorial Health Warnings among Mexican Youth and Adults: A Population-level Intervention with Potential to Reduce Tobacco-Related Inequities," *Cancer Causes & Control*, 2012, *23* (1), 57–67.

Handel, Benjamin R, "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts," *American Economic Review*, 2013, *103* (7), 2643–82.

Hastings, Justine S. and Jeffrey M. Weinstein, "Information, School Choice, and Academic Achievement: Evidence from Two Experiments," *Quarterly Journal of Economics*, 2008, *123* (4), 1373–1414.

Houde, Sebastien, "How Consumers Respond to Product Certification and the Value of Energy Information," *RAND Journal of Economics*, 2018, *49* (2), 453–477.

_ , "The Incidence of Coarse Certification: Evidence from the ENERGY STAR Program," *E2e Project Working Paper Series*, 2018. No. 036.

Hummel, Dennis and Alexander Maedche, "How Effective is Nudging? A Quantitative Review on the Effect Sizes and Limits of Empirical Nudging Studies," *Journal of Behavioral and Experimental Economics*, 2019, *80*, 47–58.

Ito, Koichiro, Takanori Ida, and Makoto Tanaka, "Moral Suasion and Economic Incentives: Field Experimental Evidence from Energy Demand," *American Economic Journal: Economic Policy*, 2018, *10* (1), 240–67.

Jensen, Robert, "The (Perceived) Returns to Education and the Demand for Schooling," *Quarterly Journal of Economics*, 2010, *125* (2), 515–548.

Kenkel, Donald S., Alan D. Mathios, Grace N. Phillips, Revathy Suryanarayana, Hua Wang, and Sen Zeng, "Fear or Knowledge: The Impact of Graphic Cigarette Warnings on Tobacco Product Choices," *NBER Working Paper*, 2023. NBER WP No. 31534.

Kiesel, Kristin and Sofia B. Villas-Boas, "Can information costs affect consumer choice? Nutritional labels in a supermarket experiment," *International Journal of Industrial Organization*, 2013, *31* (2), 153–163.

Kliemann, N., J. Wardle, F. Johnson, and H. Croker, "Reliability and Validity of a Revised Version of the General Nutrition Knowledge Questionnaire," 2016, *70* (10), 1174–1180.

Knittel, Christopher R. and James M. Sallee, "Automobile Depreciation: Estimation and Implications," 2019. Working paper, available from authors upon request.

Knittel, Christopher R and Samuel Stolper, "Using Machine Learning to Target Treatment: The Case of Household Energy Use," *NBER Working Paper*, 2019. NBER WP No. 26531.

List, John A., Matthias Rodemeier, Sutanuka Roy, and Gregory K. Sun, "Judging Nudging: Understanding the Welfare Effects of Nudges Versus Taxes," *NBER Working Paper*, 2023. NBER WP No. 31152.

Loewenstein, George and Nick Chater, "Putting Nudges in Perspective," *Behavioral Public Policy*, 2017, *1*, 26–53.

Loschel, Andreas, Matthias Rodemeier, and Madeline Werthschulte, "Can Self-Set Goals Encourage Resource Conservation? Field Experimental Evidence from a Smartphone App," *ZEW - Centre for European Economic Research Discussion Paper*, 2022. No. 20-039.

Madrian, Brigitte C and Dennis F Shea, "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior," *Quarterly Journal of Economics*, 2001, *116* (4), 1149–1187.

Milkman, Katherine L., Linnea Gandhi, Mitesh S. Patel, Heather N. Graci, Dena M. Gromet, Hung Ho, Joseph S. Kay, Timothy W. Lee, Jake Rothschild, Jonathan E. Bogard, Ilana Brody, Christopher F. Chabris, Edward Chang, Gretchen B. Chapman, Jennifer E. Dannals, Noah J. Goldstein, Amir Goren, Hal Hershfield, Alex Hirsch, Jillian Hmurovic, Samantha Horn, Dean S. Karlan, Ariella S. Kristal, Cait Lamberton, Michelle N. Meyer, Allison H. Oakes, Maurice E. Schweitzer, Maheen Shermohammed, Joachim Talloen, Caleb Warren, Ashley Whillans, Kuldeep N. Yadav, Julian J. Zlatev, Ron Berman, Chalanda N. Evans, Rahul Ladhania, Jens Ludwig, Nina Mazar, Sendhil Mullainathan, Christopher K. Snider, Jann Spiess, Eli Tsukayama, Lyle Ungar, Christophe Van den Bulte, Kevin G. Volpp, and Angela L. Duckworth, "A 680,000-person Megastudy of Nudges to Encourage Vaccination in Pharmacies," *Proceedings of the National Academy of Sciences*, 2022, *119* (6), e2115126119.

_ , Mitesh S. Patel, Linnea Gandhi, Heather N. Graci, Dena M. Gromet, Hung Ho, Joseph S. Kay, Timothy W. Lee, Modupe Akinola, John Beshears, Jonathan E. Bogard, Alison Buttenheim, Christopher F. Chabris, Gretchen B. Chapman, James J. Choi, Hengchen Dai, Craig R. Fox, Amir Goren, Matthew D. Hilchey, Jillian Hmurovic, Leslie K. John, Dean Karlan, Melanie Kim, David Laibson, Cait Lamberton, Brigitte C. Madrian, Michelle N. Meyer, Maria Modanu, Jimin Nam, Todd Rogers, Renante Rondina, Silvia Saccardo, Maheen Shermohammed, Dilip Soman, Jehan Sparks, Caleb Warren, Megan Weber, Ron Berman, Chalanda N. Evans, Christopher K. Snider, Eli Tsukayama, Christophe Van den Bulte, Kevin G. Volpp, and Angela L. Duckworth, "A Megastudy of Text-based Nudges Encouraging Patients to Get Vaccinated at an Upcoming Doctor's Appointment," *Proceedings of the National Academy of Sciences*, 2021, *118* (20), e2101165118.

Moran, Alyssa J and Christina A Roberto, "Health Warning Labels Correct Parents' Misperceptions about Sugary Drink Options," *American Journal of Preventive Medicine*, 2018, *55* (2), e19–e27.

Mullainathan, Sendhil, Joshua Schwartzstein, and William J Congdon, "A Reduced-Form Approach to Behavioral Public Finance," *Annual Review of Economics*, 2012, *4* (1), 511–540.

Nolan, Jessica M, P Wesley Schultz, Robert B Cialdini, Noah J Goldstein, and Vladas Griskevicius, "Normative Social Influence is Underdetected," *Personality and Social Psychology Bulletin*, 2008, *34* (7), 913–923.

Oak Ridge National Laboratory, "Developing a Best Estimate of Annual Vehicle Mileage for 2017 NHTS Vehicles," Technical Report undated. Available at `nhts.ornl.gov/assets/2017BESTMILE_Documentation.pdf`.

OECD (Organization for Economic Cooperation and Development), *Behavioural Insights and Public Policy: Lessons from Around the World* 2017.

Petimar, Joshua, Laura A. Gibson, Jiali Yan, Sara N. Bleich, Nandita Mitra, Marsha L. Trego, Hannah G. Lawman, and Christina A. Roberto, "Sustained Impact of the Philadelphia Beverage Tax on Beverage Prices and Sales Over 2 Years," *American Journal of Preventive Medicine*, 2022, *62* (6), 921–929.

Rodemeier, Matthias and Andreas Loschel, "Information Nudges, Subsidies, and Crowding Out of Attention: Field Evidence from Energy Efficiency Investments," *Available at SSRN 4213071*, 2022.

Sallee, James M., "The Surprising Incidence of Tax Credits for the Toyota Prius," *American Economic Journal: Economic Policy*, 2011, *3* (2), 189–219.

_ , Sarah E. West, and Wei Fan, "Do Consumers Recognize the Value of Fuel Economy? Evidence From Used Car Prices and Gasoline Price Fluctuations," *Journal of Public Economics*, 2016, *135*, 61–73.

Schultz, P Wesley, Jessica M Nolan, Robert B Cialdini, Noah J Goldstein, and Vladas Griskevicius, "The Constructive, Destructive, and Reconstructive Power of Social Norms," *Psychological Science*, 2007, *18* (5), 429–434.

Seiler, Stephan, Anna Tuchman, and Song Yao, "The Impact of Soda Taxes: Pass-Through, Tax Avoidance, and Nutritional Effects," *Journal of Marketing Research*, 2021, *58* (1), 22–49.

Silver, Lynn D., Shu Wen Ng, Suzanne Ryan-Ibarra, Lindsey Smith Taillie, Marta Induni, Donna R. Miles, Jennifer M. Poti, and Barry M. Popkin, "Changes in Prices, Sales, Consumer Spending, and Beverage Consumption One Year After a Tax on Sugar-sweetened Beverages in Berkeley, California, US: A Before-and-after Study," *PLOS Medicine*, 2017, *14* (4), 1–19.

Taillie, Lindsey Smith, Marcela Reyes, M. Arantxa Colchero, Barry Popkin, and Camila Corvalan, "An evaluation of Chile's Law of Food Labeling and Advertising on sugar-sweetened beverage purchases from 2015 to 2017: A before-and-after study," *PLoS Med*, 2020, *17* (2).

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.

Thaler, Richard H. and Cass R. Sunstein, "Libertarian Paternalism Is Not an Oxymoron," *University of Chicago Law Review*, 2003, *70* (4), 1159–1202.

Thaler, Richard H and Cass R Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Penguin, 2008.

_ and _ , *Nudge: The Final Edition*, Penguin, 2021.

Thrasher, James F, Edna Arillo-Santillán, Victor Villalobos, Rosaura Pérez-Hernández, David Hammond, Jarvis Carter, Ernesto Sebrié, Raul Sansores, and Justino Regalado-Piñeda, "Can Pictorial Warning Labels on Cigarette Packages Address Smoking-related Health Disparities? Field Experiments in Mexico to Assess Pictorial Warning Label Content," *Cancer Causes & Control*, 2012, *23* (1), 69–80.

Thunström, Linda, "Welfare Effects of Nudges: The Emotional Tax of Calorie Menu Labeling," *Judgment and Decision Making*, 2019, *14* (1), 11.

Ungemach, Christoph, Adrian R. Camilleri, Eric J. Johnson, Richkard P. Larrick, and Elke U. Weber, "Translated Attributes as Choice Architecture: Aligning Objectives and Choices Through Decision Signposts," *Management Science*, 2017, *64* (5), 1–15.

U.S. Energy Information Administration, "U.S. Average Retail Gasoline Prices in 2019 Were Slightly Lower than in 2018," Technical Report 2020. Available at www.eia.gov/todayinenergy/detail.php?id=42435.

U.S. Government Interagency Working Group on Social Cost of Greenhouse Gases, "Technical Support Document: Social Cost of Carbon, Methane, and Nitrous Oxide," Technical Report 2021. Available at www.whitehouse.gov/wp-content/uploads/2021/02/TechnicalSupportDocument_SocialCostofCarbonMethaneNitrousOxide.pdf.

Wang, Y Claire, Pamela Coxson, Yu-Ming Shen, Lee Goldman, and Kirsten Bibbins-Domingo, "A Penny-Per-Ounce Tax on Sugar-Sweetened Beverages Would Cut Health and Cost Burdens of Diabetes," *Health Affairs*, 2012, *31* (1), 199–207.

Weyl, E. Glen and Michal Fabinger, "Pass-Through as an Economic Tool: Principles of Incidence under Imperfect Competition," *Journal of Political Economy*, 2013, *121* (3), 528–583.

Yong, Peirre L., John Bertko, and Richard Kronick, "Actuarial Value and Employer-Sponsored Insurance," Technical Report 2011. Available at https://aspe.hhs.gov/reports/actuarial-value-employer-sponsored-insurance-0.

Table 1: **Average Treatment Effects, Variance, and Covariance**

(a) **Cars Experiment**

	(1) OLS	(2) OLS	(3) Mixed effects	(4) Mixed effects
Treated	-59.10**	-55.81	-58.70**	-56.69
	(23.30)	(57.44)	(23.42)	(57.42)
Bias × Treated		-0.00		-0.01
		(0.04)		(0.04)
Externality × Treated		-0.06		-0.01
		(1.06)		(1.06)
Cov(bias, treatment effect)		-1,106		-7,744
(standard error)		20,902		21,348
Cov(externality, treatment effect)		-35		-37
(standard error)		513		514
Var(treatment effect)			30,428	
(standard error)			11,925	
Number of participants	1,267	1,267	1,267	1,267
Number of observations	2,534	2,534	2,534	2,534

(b) **Drinks Experiment**

	(1) OLS	(2) OLS	(3) Mixed effects	(4) Mixed effects
Treated	-0.43***	-0.65***	-0.43***	-0.65***
	(0.04)	(0.09)	(0.04)	(0.09)
Bias × Treated		0.08**		0.08**
		(0.04)		(0.04)
Cov(bias, treatment effect)		0.130		0.129
(standard error)		0.055		0.055
Var(treatment effect)			0.743	
(standard error)			0.187	
Number of participants	2,619	2,619	2,619	2,619
Number of observations	7,857	7,857	7,857	7,857

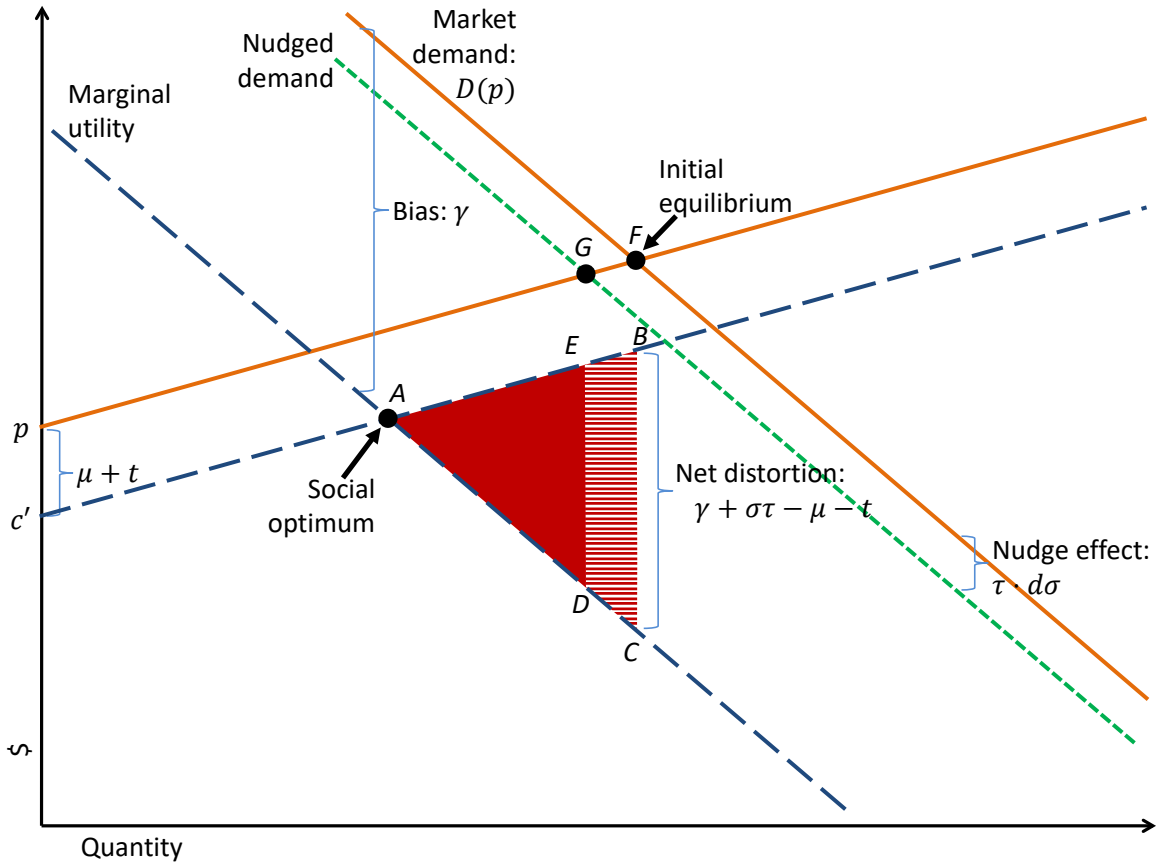
Notes: Panels (a) and (b), respectively, present estimated ATEs, variances, and covariances for the cars experiment and drinks experiment, pooling across all labels. Columns 1 and 2 present fixed coefficient (OLS) versions of equations (12) and (16), respectively, while columns 3 and 4 present the full random coefficient (mixed effects) models. All regressions also include controls for bias and externality as well as indicators for product pairs j and MPL order.

Table 2: **Parameters and Welfare Analysis: Pooled Labels**

Parameter	Description	(1) Cars experiment	(2) Drinks experiment
D'_p	Demand slope (share of purchases/($/unit)	−0.00060	−0.14
$\mathbb{E}[\gamma]$	Average bias ($/unit)	135	2.56
		(17.6)	(0.02)
$\mathbb{E}[\phi]$	Average externality ($/unit)	50	1.22
		(0.62)	(0.00)
$\mathbb{E}[\tau]$	Average treatment effect ($/unit)	−59	−0.43
		(23)	(0.04)
$Var[\tau]$	Treatment effect variance (($/unit)^2)	30,428	0.74
		(11,925)	(0.19)
$Cov[\gamma, \tau]$	Bias and treatment effect covariance (($/unit)^2)	$-7,744$	0.13
		$(21,348)$	(0.05)
$Cov[\phi, \tau]$	Externality and treatment effect covariance (($/unit)^2)	-37	0
		(514)	
ρ	Pass-through (unitless)	0.80	0.80
μ	Markup ($/unit)	0	0
$\Delta W(t=0)$	Total surplus effect with no tax ($/unit)	−0.07	0.11
	if homogeneous effect: $Cov[\delta, \tau] = Var[\tau] = 0$	4.36	0.18
	if zero average effect: $\mathbb{E}[\tau] = 0$	−4.43	−0.07
	if pure noise: $\mathbb{E}[\tau] = Cov[\delta, \tau] = 0$	−9.07	−0.05
$\Delta W(t = t^*)$	Total surplus effect with optimal tax ($/unit)	−4.43	−0.07

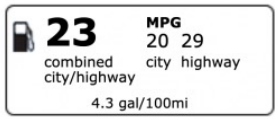
Notes: This table presents parameter estimates and total surplus effects, pooling across labels in each experiment. $\Delta W(t=0)$ and $\Delta W(t=t^*)$ are computed using equations (19) and (20), given the parameters reported above. "Unit" is "vehicle-year" for cars and "12-pack" for sugary drinks. Standard errors are in parentheses.

Figure 1: **Effect of Marginal Nudge on Total Surplus**



Notes: This figure illustrates the effect of a marginal nudge on total surplus, assuming homogeneous $\{\gamma, \tau\}$ for illustration. See text for details.

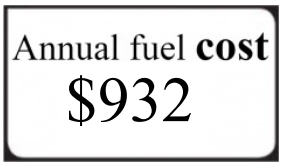
Figure 2: **Labels**



Full MPG label



Average cost label



Personalized cost label



SmartWay label

(a) **Cars Experiment**



Nutrition facts label



Stop sign warning label



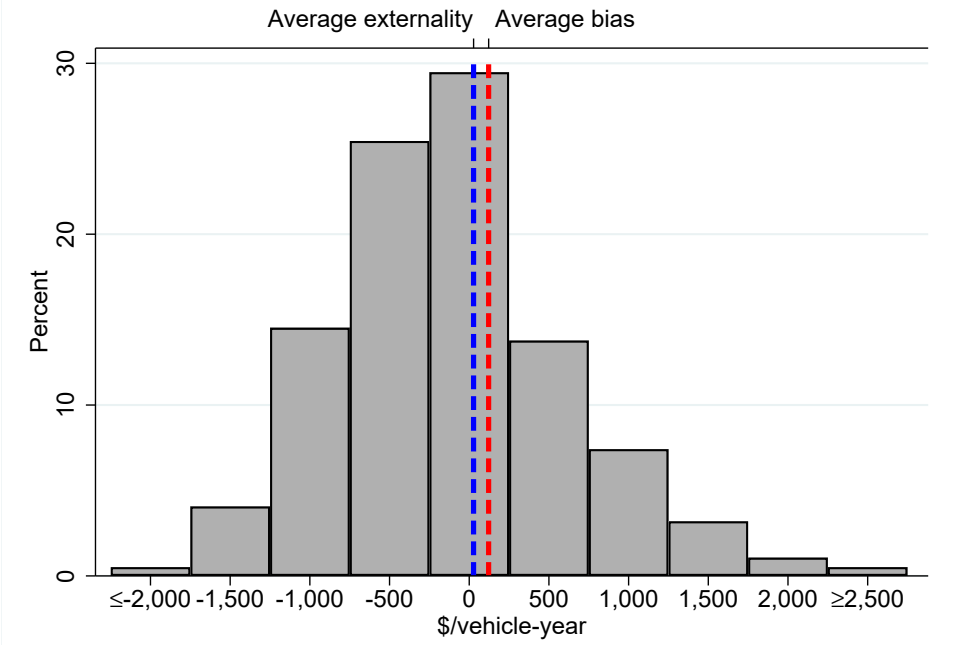
Graphic warning label

(b) **Drinks Experiment**

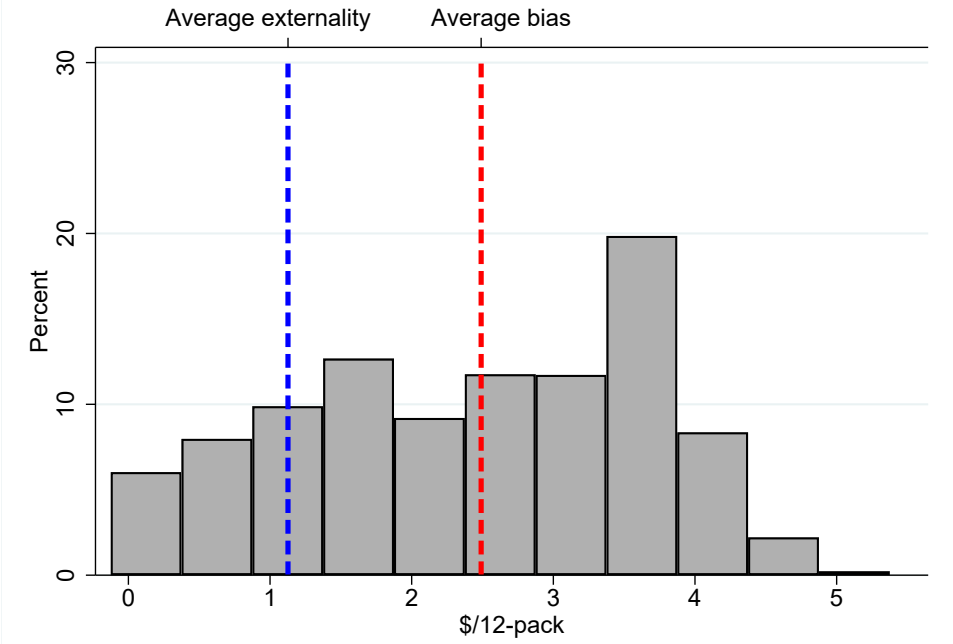
Notes: Panels (a) and (b), respectively, present the labels used in the cars experiment and drinks experiment.

Figure 3: **Distributions of Bias Estimates**

(a) **Cars Experiment**



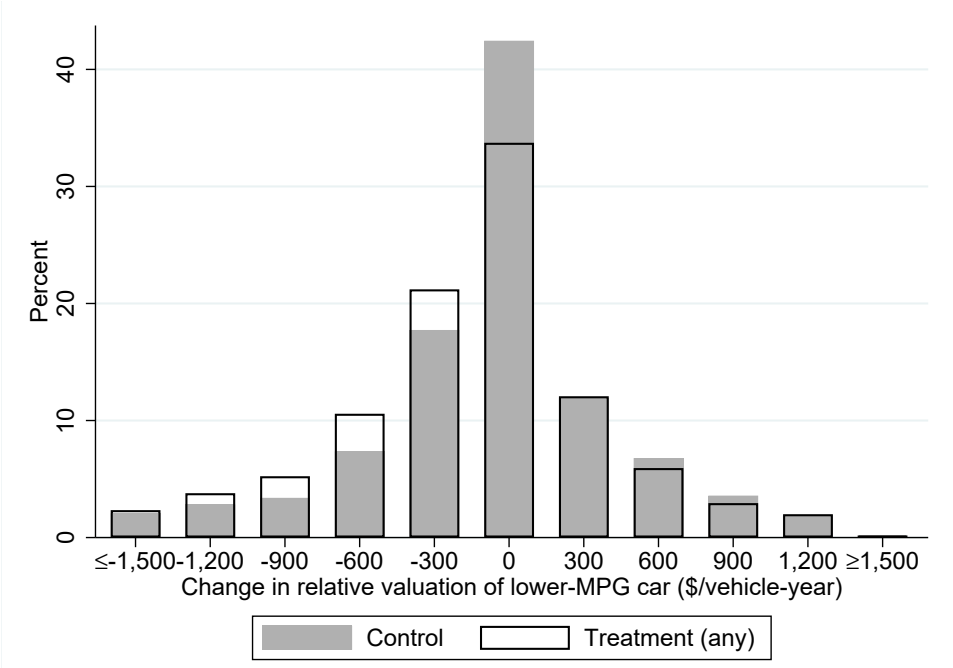
(b) **Drinks Experiment**



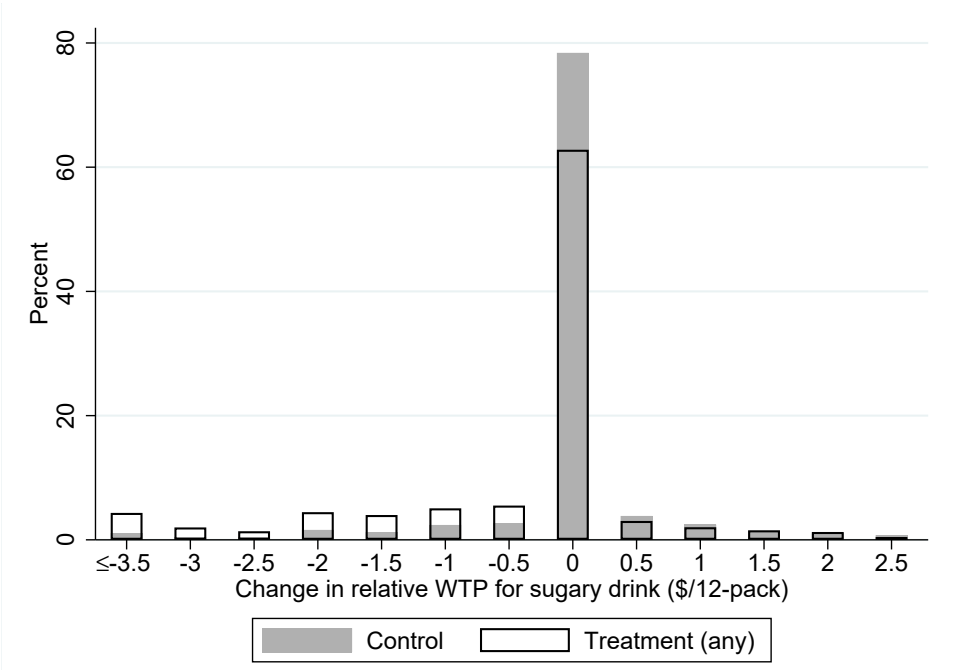
Notes: Panels (a) and (b), respectively, present the distributions of bias estimates across participants in the cars experiment and drinks experiment. The vertical lines are the average bias and externality.

Figure 4: **Changes in Willingness-to-Pay**

(a) **Cars Experiment**



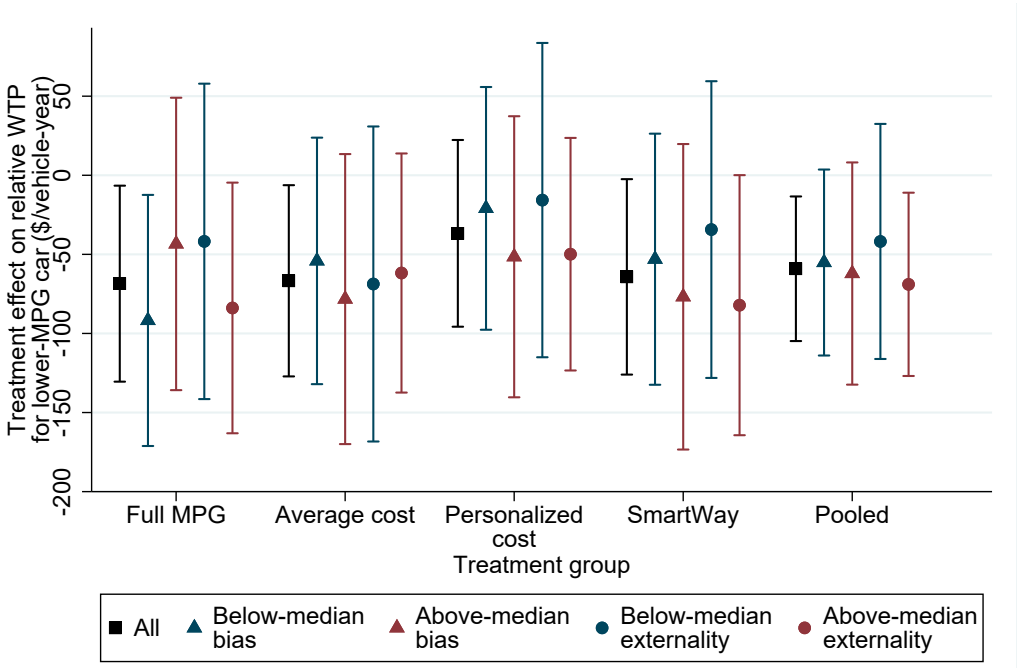
(b) **Drinks Experiment**



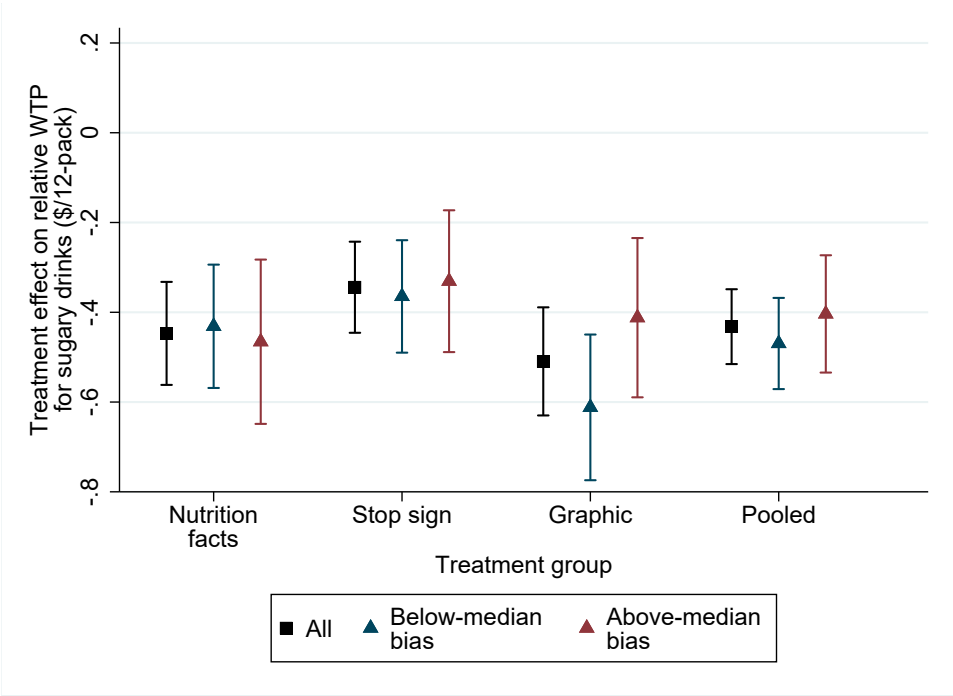
Notes: Panels (a) and (b), respectively, present histograms of baseline-endline willingness-to-pay changes in the cars experiment and drinks experiment.

Figure 5: **Average Treatment Effects and Heterogeneity**

(a) **Cars Experiment**

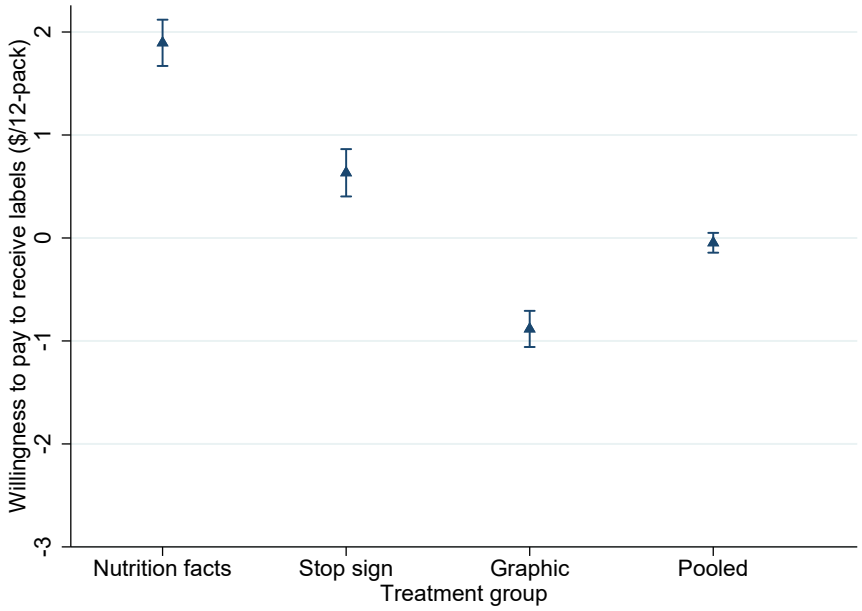


(b) **Drinks Experiment**



Notes: Panels (a) and (b), respectively, present estimates of equation (12) for the cars experiment and drinks experiment, also controlling for product pair and choice order indicators. For each label, the coefficients on the left are average treatment effects in the full samples. The remaining coefficients are heterogeneous effects in subgroups of observations with above- or below-median bias or externality estimates. In the drinks experiment, we assume homogeneous externalities, so there are no above- and below-median externality subgroups.

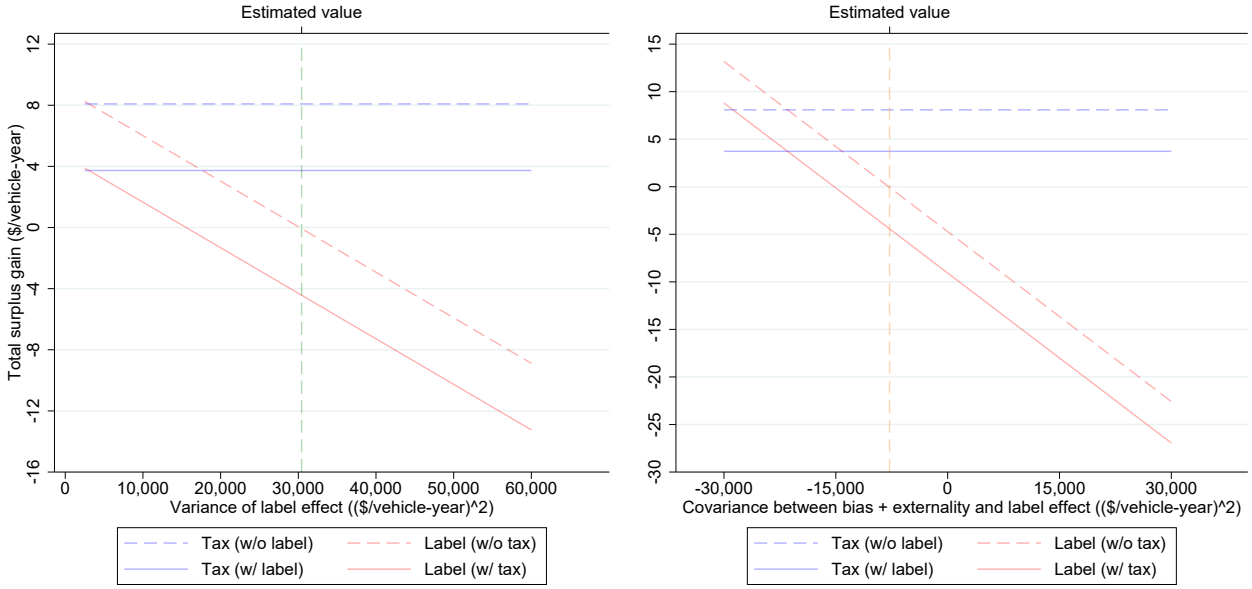
Figure 6: **Willingness-to-Pay to Receive Labels**



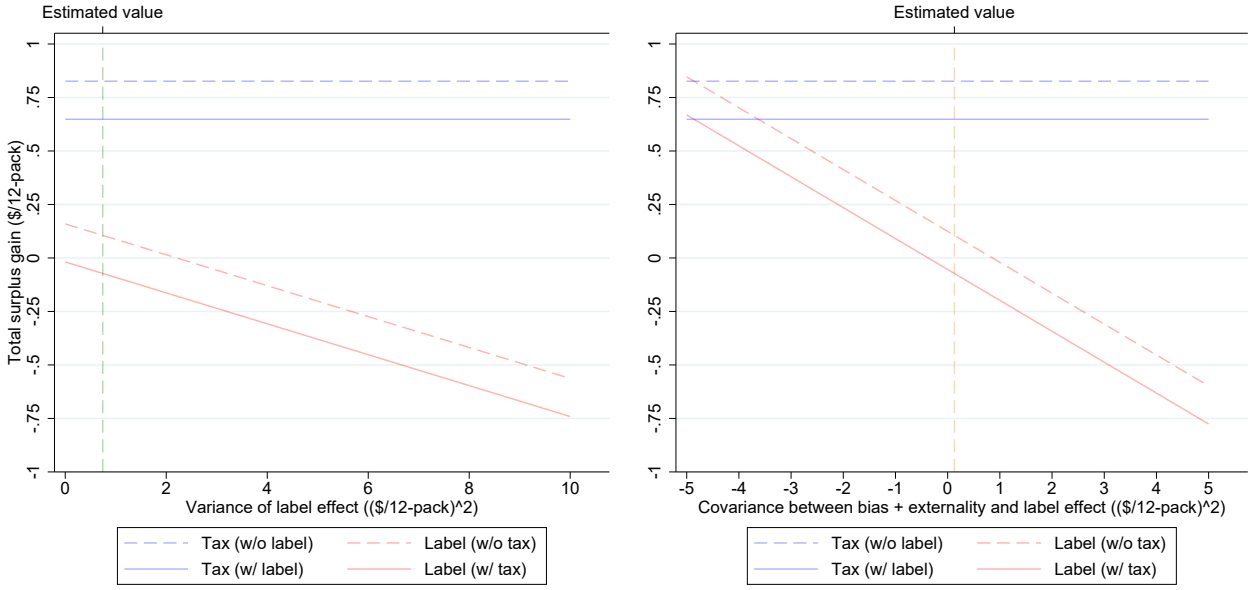
Notes: For this survey question, participants were told to assume that they had been selected to receive 12-packs of sugary drinks. The survey then elicited their hypothetical willingness-to-pay to receive or avoid receiving labels on the packages. This figure presents the average hypothetical WTP to receive labels.

Figure 7: **Total Surplus Effects Under Alternative Targeting Assumptions**

(a) **Cars Experiment**



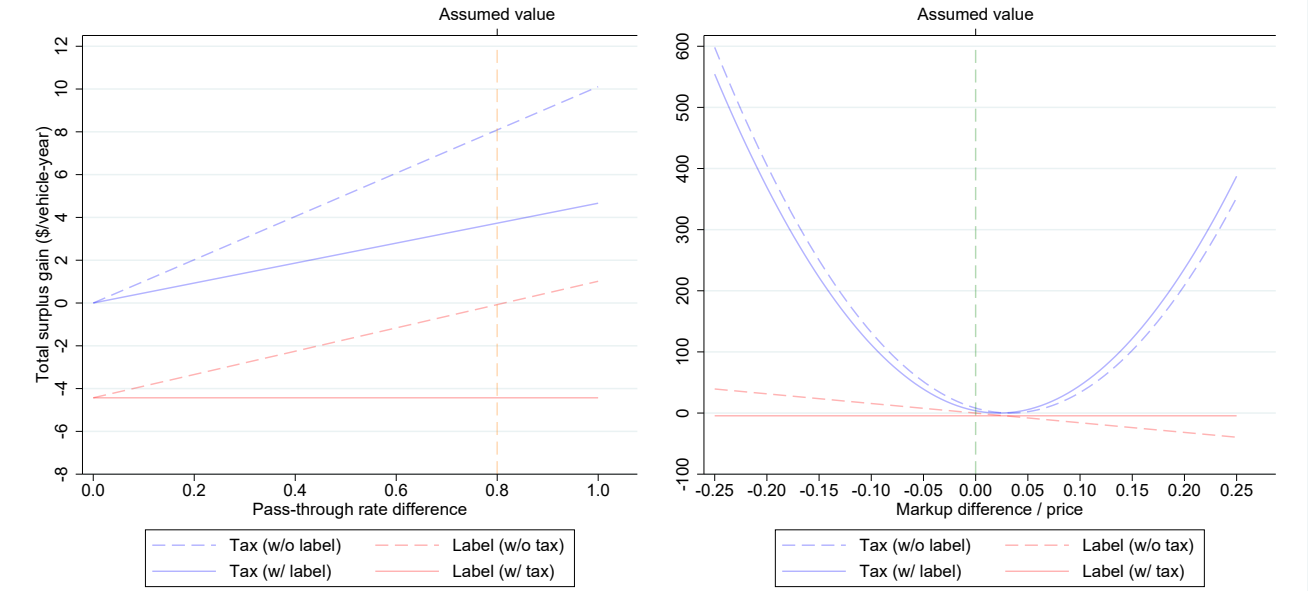
(b) **Drinks Experiment**



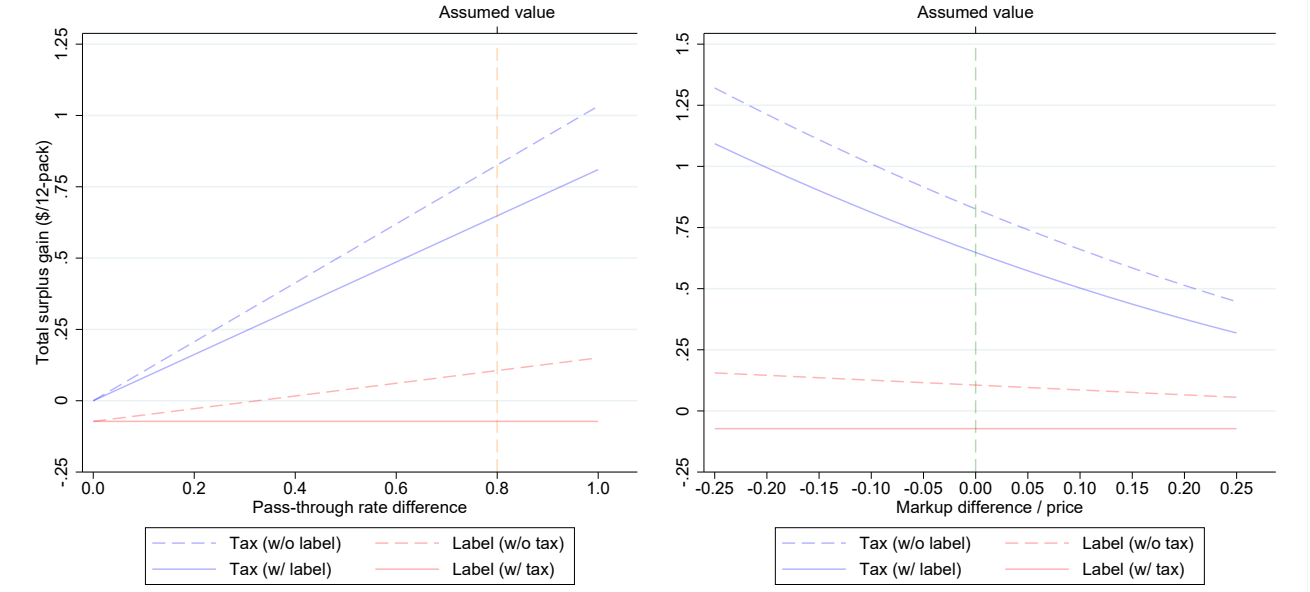
Notes: This figure presents the effects of labels and taxes on total surplus under alternative assumptions for $Var(\tau)$ and $Cov(\tau, \delta)$. The total surplus effects are computed using equations (19) and (20), given the other parameters reported in Table 2. The total surplus gain from the optimal tax $t^* = \mathbb{E}[\delta + \sigma\tau] - \mu$ is $-\frac{1}{2}t^{*2}D'_p$.

Figure 8: **Total Surplus Effects Under Alternative Pass-Through and Markup Assumptions**

(a) **Cars Experiment**



(b) **Drinks Experiment**



Notes: This figure presents the effects of labels and taxes on total surplus under alternative assumptions for ρ and μ. The total surplus effects are computed using equations (19) and (20), given the other parameters reported in Table 2. The total surplus gain from the optimal tax $t^* = \mathbb{E}[\delta + \sigma\tau] - \mu$ is $-\frac{1}{2}t^{*2}D'_p$.

Online Appendix

When Do "Nudges" Increase Welfare?

Hunt Allcott, Daniel Cohen, William Morrison, and Dmitry Taubinsky

Table of Contents

A Theory Appendix **49**
A.1 Relation to Deliberative Competence Metrics of Ambuehl, Bernheim and Lusardi (2022) . . . 49
A.2 Impacts on Consumer and Producer Surplus . . . 49
A.3 Proofs of Lemma 1) and Propositions 1 and 2 . . . 50
A.4 Generalization to Many Goods . . . 55

B Experimental Design Appendix **57**
B.1 Cars Experiment . . . 57
B.2 Drinks Experiment . . . 60

C Data Appendix **62**

D Empirical Results Appendix **65**
D.1 Estimates of Table 1 by Label . . . 69
D.2 Alternate Covariance Estimation Strategy for the Drinks Experiment . . . 74

E Welfare Analysis Appendix **75**

A Theory Appendix

A.1 Relation to Deliberative Competence Metrics of Ambuehl, Bernheim and Lusardi (2022)

Ambuehl, Bernheim and Lusardi (2022) propose measures of deliberative competence, which they use to evaluate financial literacy interventions. Ambuehl et al. evaluate frames that affect choice without directly affecting utility, which in our terminology is equivalent to nudges with $\iota \equiv 0$. Ambuehl et al. also consider situations where all distortions come from consumer bias, and not externalities. Under these assumptions, they propose the following metrics:

Definition 1. A nudge improves deliberative competence under the L1 metric if $\mathbb{E}[|\gamma+\tau|] < \mathbb{E}[|\gamma|]$ and it improves deliberative competence under the L2 metric if $\mathbb{E}[(\gamma+\tau)^2] < \mathbb{E}[\gamma^2]$.

There are several differences between our welfare metrics and these definitions. First, because Ambuehl et al. study environments with an ex-ante unknown price, their metrics apply to the full population, rather than to marginal consumers. By contrast, our welfare formulas concern markets with observed producer prices. Thus, if the nudge affects the population versus the marginal consumers differentially, there will be a fundamental disconnect between our metrics and theirs.

Of course, one can adapt their definition to marginal consumers as well to make it more comparable, and we now focus on this more comparable definition:

Definition 2. Choosing a nudge with intensity $\sigma = 1$ rather than $\sigma = 0$ improves deliberative competence under the L1 metric if $\mathbb{E}_m[\gamma+\sigma\tau]$ is decreasing in $\sigma \in [0,1]$, and it improves deliberative competence under the L2 metric if $\mathbb{E}_m[(\gamma+\sigma\tau)^2]$ is decreasing in $\sigma \in [0,1]$.

Under this definition, minimizing the L2 metric corresponds to the special case of Proposition 1 under the assumptions that markets are perfectly competitive, that the pass-through parameter is $\rho = 1$, that the tax is $t = 0$, and that there is no aversiveness. If one of those assumptions fails, Proposition 1 shows that improvements in deliberative competence don't correspond to improvements in social surplus. These various failures are illustrated in Examples 4, 5, and 6, where the nudge improves deliberative competence but does not increase social surplus. For similar reasons, minimizing the L2 metric need not correspond to increases in consumer surplus, which is formalized in Proposition 2 below. Finally, it is clear that minimizing the L1 metric need not correspond to increases in social or consumer surplus under an even larger set of assumptions.

A.2 Impacts on Consumer and Producer Surplus

Lemma 1. *Suppose that* $\mu\frac{d\varepsilon_D}{d\sigma} = -\mu\mathbb{E}_m[\tau]\frac{d\varepsilon_D}{dp}$. *The equilibrium market price* p *varies with the nudge intensity* σ *as follows:*

$$\frac{dp}{d\sigma} = (1-\rho)\mathbb{E}_m[\tau]. \tag{23}$$

The assumption of the lemma holds when $\mu = 0$ or when the demand elasticity is approximately constant in σ and p. When $\mu > 0$ and the elasticity is not approximately constant, the assumption

also mechanically holds when τ is homogeneous. Another example of when the assumption holds is when the demand curves D^τ that correspond to each set of consumers that experience a given treatment effect τ have the same elasticity.

The lemma shows that the lower is the pass-through of taxes to final consumer prices, the larger is the impact of nudges on producer prices. When $\rho < 1$, any nudge that increases demand for a product will lead to higher producer prices, and thus potentially harm all consumers, irrespective of their bias. Thus, even if a nudge stimulates demand in a socially efficient way, it might do so by transferring surplus from consumers toproducers, with the size of the transfer potentially larger than the efficiency gain itself. Conversely, nudges that increase social efficiency by depressing demand also lead to lower equilibrium prices, which generates additional benefits to consumers beyond improvements to decision quality.

It may also help to explicitly note that this effect on prices is not in some informal sense "second order" relative to the other effects of the nudges, so that the effects on prices can be argued to be negligible relative to the conjectured benefits of "light-touch" interventions that have relatively small effects on behavior. We formalize this below by quantifying the effects on consumer surplus in markets with taxes fixed at $t = 0$.

Proposition 2. *Let q^* denote the equilibrium quantity purchased in the market. With a fixed tax t, the impacts of the nudge on consumer surplus W_C and producer surplus W_P are respectively given by*

$$\frac{dW_C}{d\sigma} = \frac{1}{2}\left((1-\rho)\frac{\partial}{\partial\sigma}Var_m\left[\gamma+\sigma\tau\right] + \frac{1}{2}\rho\frac{\partial}{\partial\sigma}\mathbb{E}_m\left[(\gamma+\sigma\tau)^2\right]\right)D'_p \tag{24}$$

$$- (1-\rho)\mathbb{E}_m[\tau]q^* + \frac{\partial I}{\partial\sigma} + (1-\rho)\mathbb{E}_m[\tau]\frac{\partial I}{\partial p} \tag{25}$$

$$\frac{dW_P}{d\sigma} = (1-\rho)\mathbb{E}_m[\tau]q^* - \mu\rho\mathbb{E}_m[\tau]D'_p \tag{26}$$

For example, when $\mu = I \equiv 0$, the impact on consumer surplus can be written as $\frac{dW_C}{d\sigma} = \frac{dW}{d\sigma} - (1-\rho)\mathbb{E}_m[\tau]q^*$; i.e., the impact on social surplus minus the impact on prices. The example in Section 1.2 has shown that even nudges that only "debias" consumers can decrease social surplus. Proposition 2 thus shows that such nudges can have an even more negative effects on consumer surplus if they increase demand for the product and therefore raise prices.

A.3 Proofs of Lemma 1) and Propositions 1 and 2

This appendix presents a series of derivations that together contain the proof of Proposition 1. Proofs of Lemma 1 and Proposition 2 are intermediate results derived in Appendices A.3.1 and A.3.3, respectively.

A.3.1 Pass-Through Formula (Proof of Lemma 1)

Consider the Lerner index $\theta := \frac{p-c'(q)-t}{p}\varepsilon_D$, which we assumed to be constant. Differentiating the equation $\theta p = (p - c'(q) - t)\varepsilon_D$ with respect to σ yields

$$\theta \frac{dp}{d\sigma} = \left(\frac{dp}{d\sigma} - c''(q)\frac{dq}{d\sigma}\right)\varepsilon_D + (p - c'(q) - t)\frac{d\varepsilon_D}{d\sigma}. \tag{27}$$

Now the equilibrium demand response $\frac{dq}{d\sigma}$ is

$$\frac{dq}{d\sigma} = \frac{\partial D}{\partial \sigma} + \frac{\partial D}{\partial p}\frac{dp}{d\sigma} = -\frac{\partial D}{\partial p}\mathbb{E}_m[\tau] + \frac{\partial D}{\partial p}\frac{dp}{d\sigma}. \tag{28}$$

Plugging equation (28) into (27) thus implies that

$$\theta \frac{dp}{d\sigma} = \left(\frac{dp}{d\sigma} - c''(q)(-D'_p\mathbb{E}_m[\tau] + D'_p\frac{dp}{d\sigma})\right)\varepsilon_D + (p - c'(q) - t)\frac{d\varepsilon_D}{d\sigma}, \tag{29}$$

and thus

$$\frac{dp}{d\sigma}\left(1 - \theta - c''(q)D'_p\right) = -c''(q)D'_p\mathbb{E}_m[\tau] - (p - c'(q) - t)\frac{d\varepsilon_D}{d\sigma}, \tag{30}$$

or

$$\frac{dp}{d\sigma} = \frac{-c''(q)D'_p\mathbb{E}_m[\tau] - \mu\frac{d\varepsilon_D}{d\sigma}}{1 - \theta - c''(q)D'_p}. \tag{31}$$

Analogously, a tax t_c on consumers changes producer prices as follows (noting that in this case a tax is just a special case of a nudge with $\tau \equiv 1$):

$$\frac{dp}{dt_c} = \frac{c''(q)D'_p - \mu\frac{d\varepsilon_D}{dp}}{1 - \theta - c''(q)D'_p}. \tag{32}$$

The pass-through is $\frac{dp}{dt} = \rho = 1 + \frac{dp}{dt_c}$. Thus, if $\mu\frac{d\varepsilon_D}{d\sigma} = -\mu\mathbb{E}_m[\tau]\frac{d\varepsilon_D}{dp}$, then

$$\frac{dp}{d\sigma} = -\frac{dp}{dt}\mathbb{E}_m[\tau] = (1 - \rho)\mathbb{E}_m[\tau]. \tag{33}$$

For example, $\mu\frac{d\varepsilon_D}{d\sigma} = -\mu\mathbb{E}_m[\tau]\frac{d\varepsilon_D}{dp}$ holds with constant-elasticity demand or homogeneous treatment effects. This establishes Lemma 1.

A.3.2 Optimal Tax Formula

The tax must maximize

$$W = \int_{v \geq p(t)-\gamma-\sigma\tau} v dF - c(q^*) + I. \tag{34}$$

Differentiating yields

$$\frac{dW}{dt} = -c'(q^*)\frac{dq^*}{dt} \tag{35}$$

$$- \int_{v=p(t)-\gamma-\sigma\tau} f(p(t)-\gamma-\sigma t)(p(t)-\gamma-\sigma\tau)(p'(t)) + \frac{\partial I}{\partial p}\frac{dp}{dt} \tag{36}$$

$$= -c'(q^*)\frac{dq^*}{dt} + D_p p'(t)\left(p(t) - \mathbb{E}_m[\gamma+\sigma\tau]\right) + \rho\frac{\partial I}{\partial p} \tag{37}$$

$$= \frac{dD}{dt}\left(p(t) - \mathbb{E}_m[\gamma+\sigma\tau]\right) - c'(q^*)) + \rho\frac{\partial I}{\partial p}, \tag{38}$$

where $q^* = Pr(v \geq p(t) - \gamma - \sigma\tau)$.

Substituting $p(t) - c' = \mu + t$ implies that

$$W'(t) = \frac{dD}{dt}\left(\mu + t - \mathbb{E}_m[\gamma+\sigma\tau]\right) + \rho\frac{\partial I}{\partial p}. \tag{39}$$

Setting $W'(t^*) = 0$ thus implies that

$$t^* = \mathbb{E}_m[\gamma+\sigma\tau] - \mu - \rho\frac{\frac{dI}{dp}}{\frac{dD}{dt}}, \tag{40}$$

or alternatively,

$$t^* = \mathbb{E}_m[\gamma+\sigma\tau] - \mu - \sigma\mathbb{E}_m\left[\Delta\iota\right], \tag{41}$$

where $\sigma\mathbb{E}_m\left[\Delta\iota\right]$ is the average difference in psychic costs that consumers on the margin obtain from purchasing the good versus not.

Under the assumption that terms of order $\frac{d^2D}{dt^2}t^2$ and $\frac{d}{dt}\frac{\partial}{\partial p}Dt^2$ are negligible, the welfare impact of the optimal tax is

$$W''(t^*)t^{*2}/2 = -\frac{dD}{dt}_{t=0} t^{*2}/2. \tag{42}$$

A.3.3 Impacts on Consumer, Producer, and Social Surplus in the Absence of Taxes

Consumer surplus is given by $W_C = \int_{v\geq p-\gamma-\sigma\tau} v dF - pq^* + I$, producer surplus is given by $W_P = p - c(q^*)$, and social surplus is given by $W = \int_{v\geq p-\gamma-\sigma\tau} v dF - c(q^*) + I$.

Equation (28) and Lemma 1 imply that the impact of σ on equilibrium quantity q^* is

$$\frac{dq^*}{d\sigma} = -\frac{\partial D}{\partial p}\mathbb{E}_m[\tau] + \frac{\partial D}{\partial p}(1-\rho)\mathbb{E}_m[\tau] = -\rho\mathbb{E}_m[\tau]D'_p. \tag{43}$$

Thus

$$\frac{d}{d\sigma}c(q^*) = c'(q^*)\rho\mathbb{E}_m[\tau]D'_p. \tag{44}$$

Using the multidimensional Leibniz rule, we have that

$$\frac{d}{d\sigma}\int_{v\geq p-\gamma-\sigma\tau} vdF = -\int_{v=p-\gamma-\sigma\tau}(p-\gamma-\sigma\tau)\left(-\tau+\frac{dp}{d\sigma}\right)dF \tag{45}$$

$$= \mathbb{E}_m[(p-\gamma-\sigma\tau)(-\tau+(1-\rho)\mathbb{E}_m[\tau])]D'_p \tag{46}$$

$$= -p\rho\mathbb{E}_m[\tau]D'_p - \mathbb{E}_m[(\gamma+\sigma\tau)(-\tau+(1-\rho)\mathbb{E}_m[\tau])]D'_p \tag{47}$$

$$= -p\rho\mathbb{E}_m[\tau]D'_p + \rho\mathbb{E}_m[(\gamma+\sigma\tau)\tau]D'_p \tag{48}$$

$$+ (1-\rho)\left(\mathbb{E}_m[(\gamma+\sigma\tau)\tau] - \mathbb{E}_m[\gamma+\sigma\tau]\mathbb{E}_m[\tau]\right)D'_p. \tag{49}$$

Now observe that for each τ,

$$\frac{1}{2}\frac{d}{d\sigma}(\gamma+\sigma\tau)^2 = \gamma\tau+\sigma\tau^2 = \tau(\gamma+\sigma\tau). \tag{50}$$

Thus,

$$\frac{1}{2}\frac{\partial}{\partial\sigma}\mathbb{E}_m[(\gamma+\sigma\tau)^2] = \mathbb{E}_m[\tau(\gamma+\sigma\tau)] \tag{51}$$

and

$$\frac{1}{2}\frac{\partial}{\partial\sigma}Var_m[(\gamma+\sigma\tau)^2] = \mathbb{E}_m[(\gamma+\sigma\tau)\tau] - \mathbb{E}_m[(\gamma+\sigma\tau)]\mathbb{E}_m[\tau]. \tag{52}$$

Substituting into our derivations of $\frac{d}{d\sigma}\int_{v\geq p-\gamma-\sigma\tau} vdF$ above we have that

$$\frac{d}{d\sigma}\int_{v\geq p-\gamma-\sigma\tau} vdF = -p\rho\mathbb{E}_m[\tau]D'_p + \frac{1}{2}(1-\rho)\frac{\partial}{\partial\sigma}Var_m\left[\gamma+\sigma\tau\right]D'_p + \frac{1}{2}\rho\frac{\partial}{\partial\sigma}\mathbb{E}_m\left[(\gamma+\sigma\tau)^2\right]D'_p. \tag{53}$$

The impact on consumer surplus is thus given by

$$\frac{dW_C}{d\sigma} = \frac{1}{2}(1-\rho)\frac{\partial}{\partial\sigma}Var_m\left[\gamma+\sigma\tau\right]D'_p + \frac{1}{2}\rho\frac{\partial}{\partial\sigma}\mathbb{E}_m\left[(\gamma+\sigma\tau)^2\right]D'_p \tag{54}$$

$$- \left(\frac{dp}{d\sigma}q^* + p\frac{dq^*}{d\sigma}\right) - p\rho\mathbb{E}_m[\tau]D'_p + \frac{dI}{d\sigma} \tag{55}$$

$$= \frac{1}{2}(1-\rho)\frac{\partial}{\partial\sigma}Var_m\left[\gamma+\sigma\tau\right]D'_p + \frac{1}{2}\rho\frac{\partial}{\partial\sigma}\mathbb{E}_m\left[(\gamma+\sigma\tau)^2\right]D'_p - (1-\rho)\mathbb{E}_m[\tau]q^* + \frac{dI}{d\sigma}. \tag{56}$$

The impact on producer surplus is given by

$$\frac{dW_P}{d\sigma} = \frac{dp}{d\sigma}q^* + p\frac{dq^*}{d\sigma} - \frac{d}{d\sigma}c(q^*) \tag{57}$$

$$= (1-\rho)\mathbb{E}_m[\tau]q^* - p\rho\mathbb{E}_m[\tau]D'_p - c'(q^*)\rho\mathbb{E}_m[\tau]D'_p \tag{58}$$

$$= (1-\rho)\mathbb{E}_m[\tau]q^* - (p-c'(q^*))\rho\mathbb{E}_m[\tau]D'_p \tag{59}$$

$$= (1-\rho)\mathbb{E}_m[\tau]q^* - \mu\rho\mathbb{E}_m[\tau]D'_p. \tag{60}$$

Putting this together, the impact on social surplus $W = W_C + W_P$ is

$$\frac{dW}{d\sigma} = \frac{1}{2}(1-\rho)\frac{\partial}{\partial\sigma}Var_m\left[\gamma+\sigma\tau\right]D_p' + \frac{1}{2}\rho\frac{\partial}{\partial\sigma}\mathbb{E}_m\left[(\gamma+\sigma\tau-\mu)^2\right]D_p' + \frac{dI}{d\sigma}. \tag{61}$$

Finally, to obtain the statement of Proposition 1, note that

$$\begin{aligned}\frac{dI}{d\sigma} &= \frac{\partial I}{\partial\sigma} + \frac{dp}{d\sigma}\frac{\partial I}{dp} \\ &= \frac{\partial I}{\partial\sigma} + (1-\rho)\mathbb{E}_m[\tau]\frac{\partial I}{\partial p}.\end{aligned} \tag{62}$$

A.3.4 Impacts on Consumer and Producer Surplus with a Fixed Tax

Our formulas for $\frac{dW_P}{d\sigma}$ and $\frac{dW_C}{d\sigma}$ are identical if there is instead a fixed tax t on producers. The reason is that on the producer side, the tax t can be considered to simply be part of the cost function, in which case all calculations are identical. On the consumer side, the tax t on producers does not independently affect consumers, given a producer price p.

A.3.5 Impacts on Social Surplus with a Fixed Tax

A fixed tax t on producers has the same social welfare effect as a tax $t_c = t$ on consumers. Now imposing a fixed tax t_c on consumers is equivalent to assuming that bias is given by $\gamma' = \gamma - t_c$.

From above, we thus trivially have that when the tax is fixed at some value t,

$$\frac{dW}{d\sigma} = \frac{1}{2}(1-\rho)\frac{\partial}{\partial\sigma}Var_m\left[\gamma+\sigma\tau\right]D_p' + \frac{1}{2}\rho\frac{\partial}{\partial\sigma}\mathbb{E}_m\left[(\gamma+\sigma\tau-t-\mu)^2\right]D_p' \tag{63}$$

where $\mu = p - c'(q) - t$, and where we use that $Var_m\left[\gamma+\sigma\tau\right] = Var_m\left[\gamma+\sigma\tau-t\right]$.

A.3.6 Impact on Social Surplus with Optimal Tax

By the envelope theorem, $\frac{dW}{d\sigma} = \frac{\partial W}{\partial\sigma}$, where the partial derivative treats the optimal tax $t^* = \mathbb{E}_m[\gamma+\sigma\tau] - \mu - \mathbb{E}_m[\Delta\iota]$ as fixed.

Now at the optimal tax,

$$\rho\mathbb{E}_m[(\gamma+\sigma\tau-t^*-\mu)\tau] = \rho\mathbb{E}[(\gamma+\sigma\tau-\mathbb{E}[\gamma+\sigma\tau])\tau] + \rho\mathbb{E}_m[\tau]\mathbb{E}_m[\Delta\iota] \tag{64}$$

$$= \frac{1}{2}\rho\frac{\partial}{\partial\sigma}Var_m\left[\gamma+\sigma\tau\right] + \rho\mathbb{E}_m[\tau]\mathbb{E}_m[\sigma\Delta\iota]. \tag{65}$$

It thus follows that at the optimal tax,

$$\frac{dW}{d\sigma} = \frac{\partial}{\partial\sigma}Var_m\left[\gamma+\sigma\tau\right]D_p' + \rho\mathbb{E}_m[\tau]\frac{\partial I}{\partial p} + \frac{dI}{d\sigma} \tag{66}$$

where we use that $\frac{\partial I}{\partial p} = \mathbb{E}_m[\sigma\Delta\iota]D'_p$. Substituting the expression in equation (62) gives

$$\frac{dW}{d\sigma} = \frac{\partial}{\partial\sigma}Var_m\left[\gamma + \sigma\tau\right]D'_p + \frac{\partial I}{\partial\sigma} + \mathbb{E}_m[\tau]\frac{\partial I}{\partial p}. \tag{67}$$

A.4 Generalization to Many Goods

More generally, suppose that there are J different types of products, indexed by $j = \{1, \ldots, J\}$, and that each consumer must buy at least one of the products. The model in the body of the paper corresponds to the special case with two products, where product $j = 2$ is an outside good with a fixed price.

A given consumer's set of valuations and biases is given by the vectors $v = (v_1, \ldots, v_j)$ and $\gamma = (\gamma_1, \ldots, \gamma_J)$. For simplicity, we assume that the nudge only directly affects valuations of a single product, which we label product 1 without loss of generality, and we let $\sigma\tau$ continue denoting the distribution of treatment effects on this product, where σ is the strength of the nudge. The demand curve for product j is $D_j(p, \sigma)$, where p is the vector of prices. Denote the own-price elasticity of demand for product j by $\varepsilon^j_D = -\frac{\partial D_j}{dp_j} \cdot \frac{p_j}{D_j}$, and denote the elasticity of demand of product i with respect to the price p_j of product j by ε^{ij}_D. The general case where the nudge affects multiple goods is an immediate corollary that is obtained by taking the sum of the effects on each good.

Each firm produces only one type of product, at cost $c_j(q)$ to for q units, and pays tax t_j per unit. Let $\theta_j := \frac{p - c'_j(q) - t_j}{p}\varepsilon^j_D$ denote the market conduct parameter for product j, and assume that it is constant. We let $\mu_j = p - c'_j(q) - t_j$ denote the markup.

We define ρ_{jk} to be the impact of a tax on producers of j on the price of product k. We denote by $\Delta p_{1j} = p_1 - p_j$ the relative price of product 1 to product j, and we let $\Delta\rho_{1j} := \rho_{11} - \rho_{1j}$ denote the pass through of t_1 to Δp_{1j}.

For any function $X(v, \gamma, \sigma, \tau)$ we define $\mathbb{E}_{ij}[X(v, \gamma, \sigma, \tau)]$ to be the conditional expectation of X over the set of consumers who are on the margin of buying either product i or j. We define $\mathbb{E}_1[X] = \sum_j \mathbb{E}_{1j}[X]$ as the expectation over the set of consumers who are on the margin for buying product 1 versus any other product. We utilize analogous notation for the covariance and variance operators. With a slight abuse of notation, we define

$$\frac{\partial}{\partial\sigma}\mathbb{E}_{1j}[X(v, \gamma, \sigma, \tau)] := \frac{d}{d\sigma'}\int_{\{(v,\gamma,\tau)|v_1 - p_1 - \sigma\tau = v_2 - p_2\}} X(v, \gamma, \sigma', \tau)dF\,|_{\sigma'=\sigma}\,. \tag{68}$$

for any function X.

A.4.1 Impact of Nudge on prices

Analogous to Lemma 1,

$$\frac{\partial D_j}{\partial\sigma} = -\frac{\partial D_j}{\partial p_1}\mathbb{E}_1[\tau]. \tag{69}$$

Similarly, consider a consumer tax t_c on product 1. The impact of a marginal change in σ on prices is equivalent to a marginal change of $\mathbb{E}_1[\tau]$ in t_c. Thus, since $\rho_{jj} = 1 + \frac{dp_j}{dt_c}$, we have that

$$\frac{dp_1}{d\sigma} = -\frac{dp_1}{dt_c}\mathbb{E}_1[\tau] = (1-\rho_{11})\mathbb{E}_1[\tau], \tag{70}$$

and more generally

$$\frac{dp_j}{d\sigma} = -\frac{dp_j}{dt_c}\mathbb{E}_{1j}[\tau] = (1-\rho_{1j})\mathbb{E}_{1j}[\tau]. \tag{71}$$

A.4.2 Impact of Nudge on Welfare

Calculations analogous to the proof of Proposition 1 imply the following:

Proposition 3. *Assume that $\frac{d}{dp_k}\varepsilon_D^{ij}$ and $\frac{d}{d\sigma}\varepsilon_D^{1j}$ are negligible in the case where $\mu > 0$ for all i, j, k. Define $\Delta\gamma_{1j} = \gamma_1 - \gamma_j$ and $\Delta\mu_{1j} = \mu_1 - \mu_j$. Then the marginal change in social surplus from a nudge in a market with taxes t_j on products j is*

$$\frac{dW}{d\sigma} = -\sum_j \left[\frac{1}{2}(1-\Delta\rho_{1j})\frac{\partial}{\partial\sigma}Var_{1j}\left[\Delta\gamma_{1j}+\sigma\tau\right] + \frac{1}{2}(\Delta\rho_{1j})\frac{\partial}{\partial\sigma}\mathbb{E}_{1j}\left[(\Delta\gamma_{1j}+\sigma\tau-t_1+t_j-\Delta\mu_{1j})^2\right]\right]\frac{\partial}{\partial p_1}D_j \tag{72}$$

$$+\frac{\partial I}{\partial\sigma} + (1-\rho_{11})\mathbb{E}_1[\tau]\frac{\partial I}{\partial p_1}. \tag{73}$$

For intuition, note that when there are only two goods, the general expression above reduces to an expression almost identical to the one in the body of the paper. The main difference is that when the price of the outside good is exogenous, the key parameter is the pass-through of the tax on good 1 to the relative price, $p_1 - p_2$, of good 1. The key bias statistic is how much people overvalue good 1 relative to good 2, $\gamma_1 - \gamma_2$. And the interaction with market power is now captured by the difference $\mu_1 - \mu_j$. The general formula for many goods is obtained by taking the sum of the welfare impacts corresponding to each pair of good 1 and some other good j.

B Experimental Design Appendix

B.1 Cars Experiment

Figure A1: **Cars Experiment: Valuation if Gas is Free**

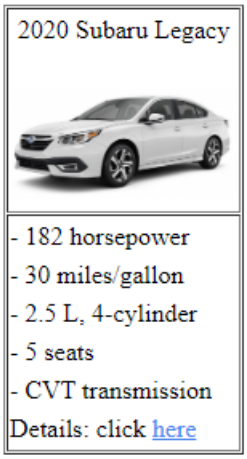


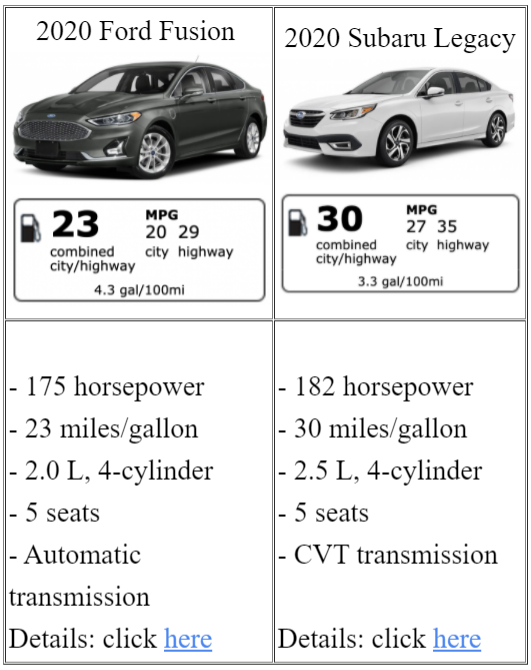
Figure A2: **Cars Experiment: Baseline Multiple Price List**

2020 Ford Fusion	2020 Subaru Legacy
	
- 175 horsepower - 23 miles/gallon - 2.0 L, 4-cylinder - 5 seats - Automatic transmission Details: click here	- 182 horsepower - 30 miles/gallon - 2.5 L, 4-cylinder - 5 seats - CVT transmission Details: click here

Please click on the choice you would prefer given the annual lease prices below.

Ford Fusion for $2000	Subaru Legacy for $2000

Figure A3: **Cars Experiment: Endline Multiple Price List with Full MPG Label**



B.2 Drinks Experiment

Figure A4: **Drinks Experiment: Recruitment Ad**



Figure A5: **Drinks Experiment: Baseline Multiple Price List**

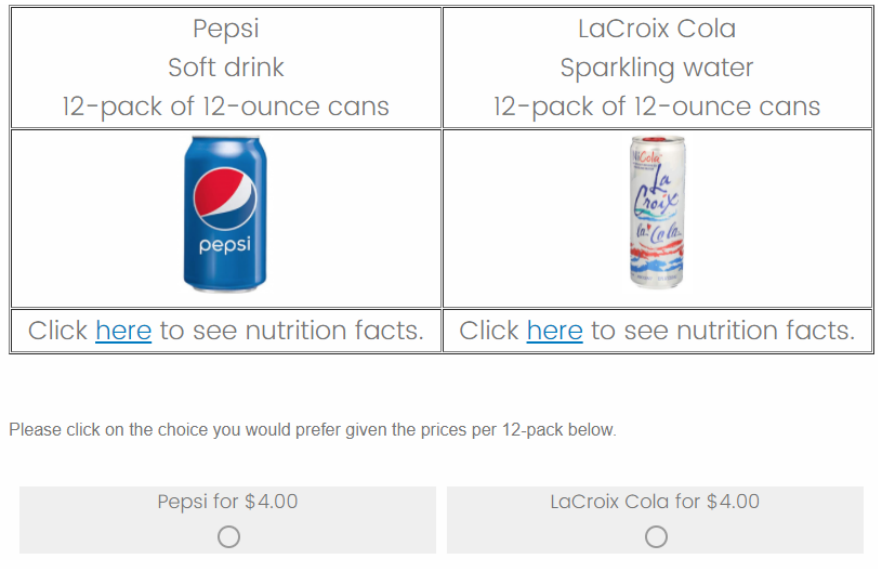
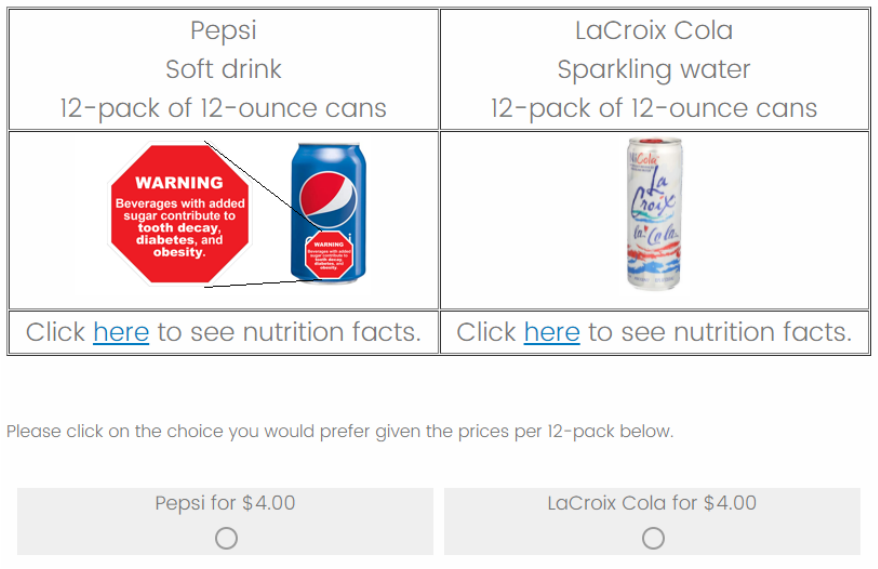


Figure A6: **Drinks Experiment: Endline Multiple Price List with Stop Sign Label**



C Data Appendix

Table A1: **Cars Experiment Descriptive Statistics**

	(1) Experiment sample	(2) US population
Income under $50,000	0.37	0.39
College degree (for age $\geq$ 25)	0.41	0.33
Male	0.53	0.49
White	0.70	0.75
Under age 45	0.41	0.44
2019 miles driven	10,803	11,131
2019 gas price ($/gallon)	2.79	2.60
Average WTP if gas is free ($/vehicle-year)	2,771	
Average baseline WTP ($/vehicle-year)	1,553	

Notes: US population averages for demographic variables are from the 2016–2020 American Community Surveys. US population average *2019 miles driven* and *2019 gas price* are from the 2017 National Household Travel Survey (Oak Ridge National Laboratory undated) and U.S. Energy Information Administration (2020), respectively. *Average WTP if gas is free* is the respondent's average valuation of the Accord, Altima, Fusion, and Legacy in the baseline questions when told to imagine that gas is free. The experiment sample includes 1,267 participants.

Table A2: **Drinks Experiment Descriptive Statistics**

	(1) Experiment sample	(2) US population
Income under $50,000	0.63	0.39
College degree (for age $\geq$ 25)	0.41	0.33
Male	0.47	0.49
White	0.84	0.75
Under age 45	0.40	0.44
Nutrition knowledge	0.70	0.70
Self-control	0.41	0.77

Notes: US population averages for demographic variables are from the 2016–2020 American Community Surveys. *Nutrition knowledge* is the share correct out of 28 questions from the General Nutrition Knowledge Questionnaire (Kliemann et al. 2016). *Self-control* is level of agreement with the statement, "I drink soda pop or other sugar-sweetened beverages more often than I should." Responses were coded as "Definitely" = 0, "Mostly" = 1/3, "Somewhat" = 2/3, and "Not at all" = 1. National averages are as reported in Allcott, Lockwood and Taubinsky (2019a). The experiment sample includes 2,619 participants.

Table A3: **Cars Experiment: Covariate Balance**

Variable	(1) Control Mean/SD	(2) Full MPG Mean/SD	(3) Fuel cost Mean/SD	(4) Personalized fuel cost Mean/SD	(5) SmartWay Mean/SD	T-test P-value (1)-(2)	(1)-(3)	(1)-(4)	(1)-(5)
Household income ($000s)	74.53	75.80	76.25	72.86	74.01	0.67	0.56	0.58	0.86
	(47.81)	(46.87)	(48.68)	(48.11)	(48.21)				
College degree	0.40	0.42	0.34	0.41	0.45	0.67	0.05**	0.72	0.13
	(0.49)	(0.49)	(0.48)	(0.49)	(0.50)				
Male	0.54	0.52	0.56	0.51	0.53	0.57	0.40	0.38	0.95
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)				
White	0.72	0.69	0.73	0.63	0.73	0.32	0.57	0.00***	0.77
	(0.45)	(0.46)	(0.44)	(0.48)	(0.45)				
Age	50.05	49.78	50.86	48.22	50.46	0.78	0.42	0.07*	0.68
	(16.42)	(15.45)	(16.38)	(16.25)	(15.95)				
N	530	494	516	492	502				
F-test of joint significance (p-value)						0.87	0.19	0.05*	0.67
F-test, number of observations						1024	1046	1022	1032

Notes: This table presents tests of covariate balance between treatment conditions in the cars experiment. The first five columns present means and standard deviations. The final four columns present p-values of t-tests of equality between each treatment condition and the control group.

Table A4: **Drinks Experiment: Covariate Balance**

Variable	(1) Control Mean/SD	(2) Nutrition Mean/SD	(3) Stop sign Mean/SD	(4) Graphic Mean/SD	T-test P-value (1)-(2)	(1)-(3)	(1)-(4)
Household income ($000s)	47.10	47.73	47.53	47.14	0.61	0.72	0.97
	(38.60)	(39.35)	(38.46)	(38.70)			
College degree	0.41	0.41	0.40	0.41	0.80	0.69	0.88
	(0.49)	(0.49)	(0.49)	(0.49)			
Male	0.47	0.47	0.45	0.48	0.87	0.13	0.73
	(0.50)	(0.50)	(0.50)	(0.50)			
White	0.84	0.85	0.83	0.84	0.50	0.29	0.95
	(0.36)	(0.36)	(0.37)	(0.36)			
Age	48.97	48.06	47.52	48.51	0.09*	0.01***	0.38
	(16.59)	(16.82)	(16.49)	(16.57)			
N	2001	1923	1980	1953			
F-test of joint significance (p-value)					0.45	0.02**	0.97
F-test, number of observations					3924	3981	3954

Notes: This table presents tests of covariate balance between treatment conditions in the drinks experiment. The first four columns present means and standard deviations. The final three columns present p-values of t-tests of equality between each treatment condition and the control group.

D Empirical Results Appendix

Table A5: **Average Treatment Effects, Variance, and Covariance for "Marginal" Consumers**

(a) **Cars Experiment**

	(1) OLS	(2) OLS	(3) Mixed effects	(4) Mixed effects
Treated	-85.18***	-82.04	-85.01***	-82.08
	(32.91)	(77.30)	(32.78)	(77.04)
Bias × Treated		-0.08		-0.09
		(0.06)		(0.06)
Externality × Treated		0.43		0.47
		(1.36)		(1.35)
Cov(bias, treatment effect)		-24,083		-26,085
(standard error)		18,988		19,177
Cov(externality, treatment effect)		-11		-7
(standard error)		652		650
Var(treatment effect)			33,254	
(standard error)			18,691	
Number of participants	482	482	482	482
Number of observations	964	964	964	964

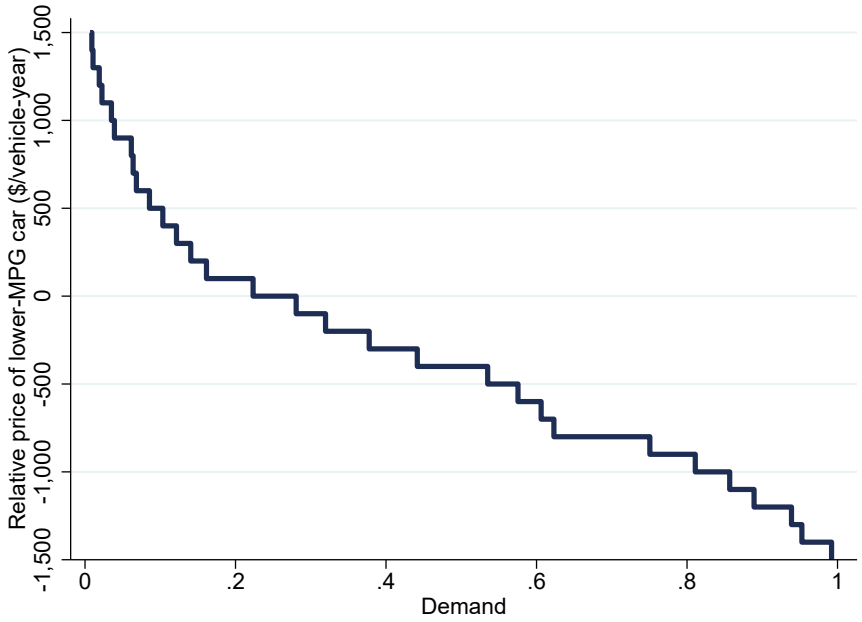
(b) **Drinks Experiment**

	(1) OLS	(2) OLS	(3) Mixed effects	(4) Mixed effects
Treated	-0.42***	-0.49***	-0.42***	-0.49***
	(0.05)	(0.10)	(0.05)	(0.10)
Bias × Treated		0.03		0.03
		(0.04)		(0.04)
Cov(bias, treatment effect)		0.043		0.043
(standard error)		0.060		0.060
Var(treatment effect)			0.666	
(standard error)			0.122	
Number of participants	983	983	983	983
Number of observations	2,949	2,949	2,949	2,949

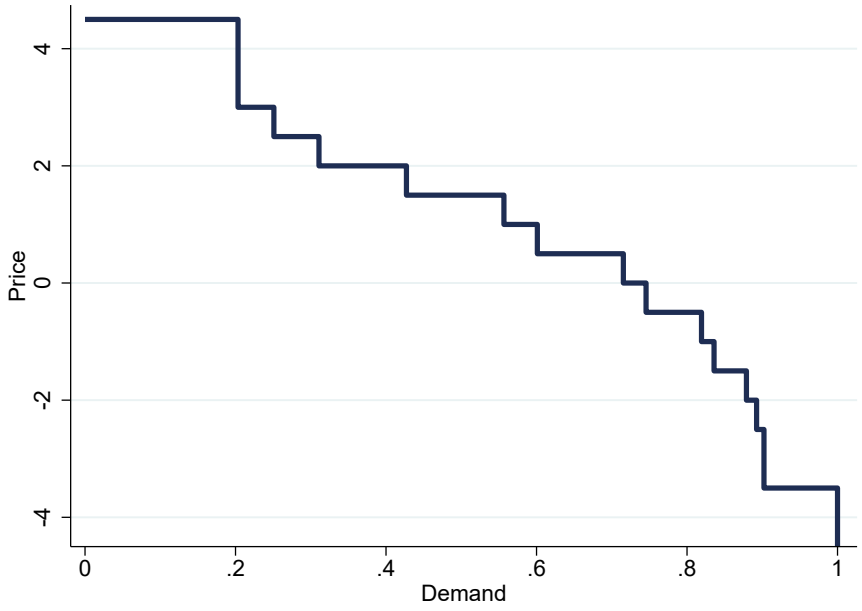
Notes: Panels (a) and (b), respectively, present estimated ATEs, variances, and covariances for the cars experiment and drinks experiment, pooling across all labels. Columns 1 and 2 present fixed coefficient (OLS) versions of equations (12) and (16), respectively, while columns 3 and 4 present the full random coefficient (mixed effects) models. The samples are limited to participants with below-median absolute value of relative WTP for both product pairs.

Figure A7: **Baseline Demand Curves**

(a) **Cars Experiment**



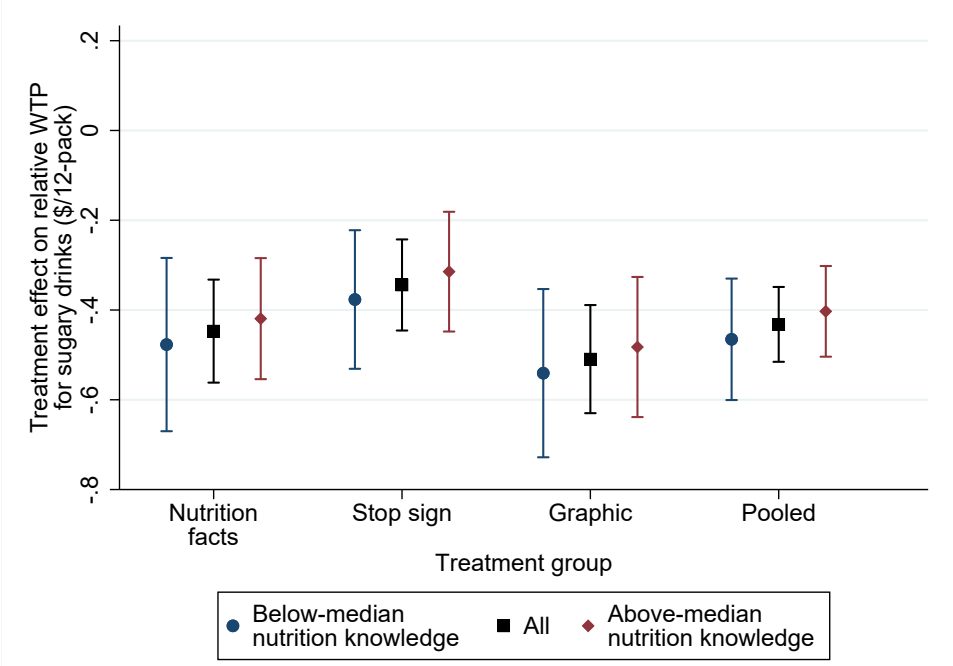
(b) **Drinks Experiment**



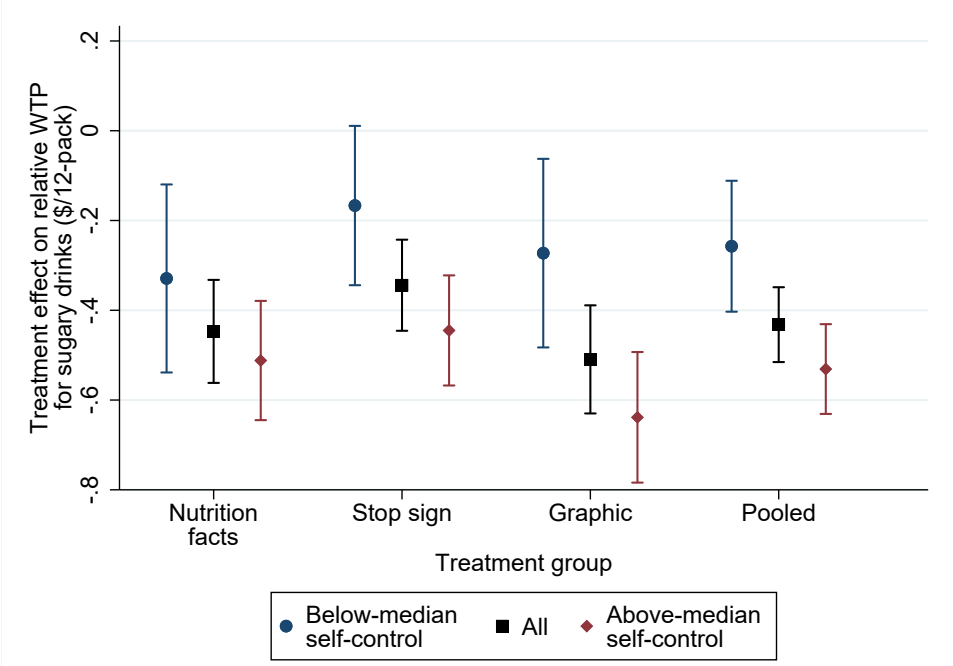
Notes: Panels (a) and (b), respectively, present the baseline demand curves for the cars experiment and drinks experiment.

Figure A8: **Drinks Experiment: Treatment Effect Heterogeneity by Bias Proxies**

(a) **Nutrition Knowledge**



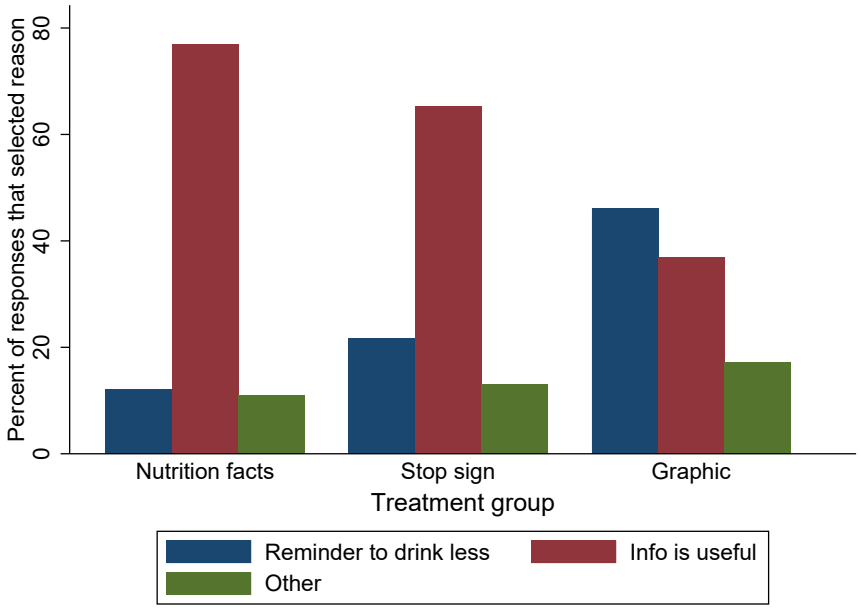
(b) **Self-Control**



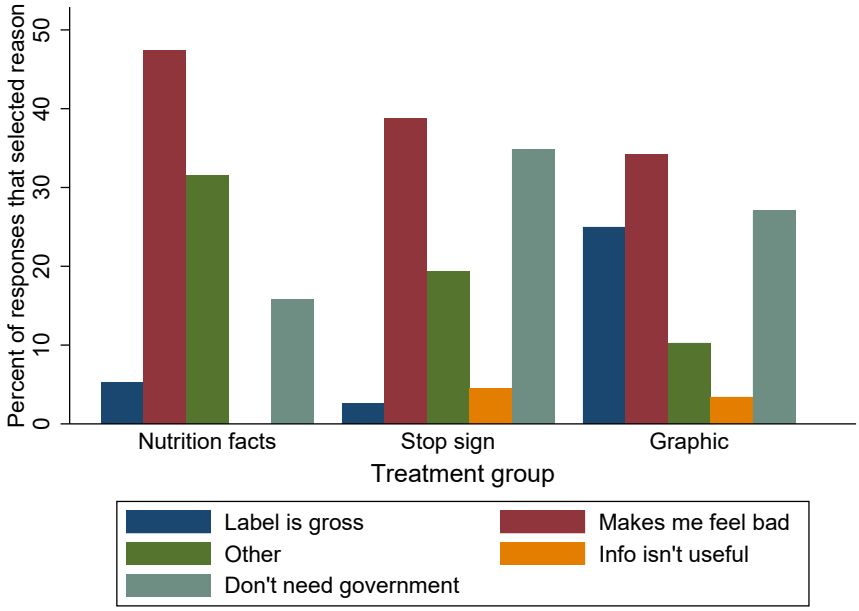
Notes: Panels (a) and (b), respectively, present estimates of equation (12) for the drinks experiment, for subgroups with above- versus below-median nutrition knowledge and self-control.

Figure A9: **Reasons for Wanting or Not Wanting Sugary Drink Labels**

(a) **Reasons for Wanting Labels**



(b) **Reasons for Not Wanting Labels**



Notes: For this survey question, participants were told to assume that they had been selected to receive 12-packs of sugary drinks. The survey then asked if they would prefer to receive drink containers with or without the label shown to their treatment group. Panels (a) and (b), respectively, present the distribution of responses to questions about why participants wanted to receive drink containers with and without the labels.

D.1 Estimates of Table 1 by Label

Table A6: **Average Treatment Effects, Variance, and Covariance for Full MPG Label (Cars Experiment)**

	(1) OLS	(2) OLS	(3) Mixed effects	(4) Mixed effects
Treated	-68.52**	-46.45	-67.85**	-48.22
	(31.53)	(77.46)	(31.57)	(77.48)
Bias × Treated		0.03		0.02
		(0.05)		(0.05)
Externality × Treated		-0.54		-0.46
		(1.46)		(1.46)
Cov(bias, treatment effect)		17,438		12,019
(standard error)		27,033		27,883
Cov(externality, treatment effect)		-176		-165
(standard error)		691		691
Var(treatment effect)			40,201	
(standard error)			17,339	
Number of participants	512	512	512	512
Number of observations	1,024	1,024	1,024	1,024

Notes: Table presents estimated ATEs, variances, and covariances for the Full MPG label in the cars experiment. Columns 1 and 2 present fixed coefficient (OLS) versions of equations (12) and (16), respectively, while columns 3 and 4 present the full random coefficient (mixed effects) models. All regressions also include controls for bias and externality as well as indicators for product pairs j and MPL order.

Table A7: **Average Treatment Effects, Variance, and Covariance for Average Cost Label (Cars Experiment)**

	(1) OLS	(2) OLS	(3) Mixed effects	(4) Mixed effects
Treated	-66.72**	-68.19	-66.16**	-66.70
	(30.77)	(77.22)	(30.93)	(77.31)
Bias × Treated		-0.01		-0.03
		(0.05)		(0.05)
Externality × Treated		0.06		0.10
		(1.41)		(1.41)
Cov(bias, treatment effect)		-6,702		-18,247
(standard error)		27,589		28,034
Cov(externality, treatment effect)		6		-19
(standard error)		681		682
Var(treatment effect)			30,179	
(standard error)			17,155	
Number of participants	523	523	523	523
Number of observations	1,046	1,046	1,046	1,046

Notes: Table presents estimated ATEs, variances, and covariances for the Average Cost label in the cars experiment. Columns 1 and 2 present fixed coefficient (OLS) versions of equations (12) and (16), respectively, while columns 3 and 4 present the full random coefficient (mixed effects) models. All regressions also include controls for bias and externality as well as indicators for product pairs j and MPL order.

Table A8: **Average Treatment Effects, Variance, and Covariance for Personalized Cost Label (Cars Experiment)**

	(1) OLS	(2) OLS	(3) Mixed effects	(4) Mixed effects
Treated	-36.72	-18.67	-35.92	-16.92
	(30.04)	(76.28)	(30.07)	(75.99)
Bias × Treated		-0.00		-0.01
		(0.05)		(0.05)
Externality × Treated		-0.35		-0.35
		(1.43)		(1.42)
Cov(bias, treatment effect)		-3,308		-7,453
(standard error)		28,016		28,448
Cov(externality, treatment effect)		-171		-184
(standard error)		663		662
Var(treatment effect)			13,550	
(standard error)			15,610	
Number of participants	511	511	511	511
Number of observations	1,022	1,022	1,022	1,022

Notes: Table presents estimated ATEs, variances, and covariances for the Personalized Cost label in the cars experiment. Columns 1 and 2 present fixed coefficient (OLS) versions of equations (12) and (16), respectively, while columns 3 and 4 present the full random coefficient (mixed effects) models. All regressions also include controls for bias and externality as well as indicators for product pairs j and MPL order.

Table A9: **Average Treatment Effects, Variance, and Covariance for SmartWay Label (Cars Experiment)**

	(1) OLS	(2) OLS	(3) Mixed effects	(4) Mixed effects
Treated	-64.23**	-85.65	-64.43**	-89.12
	(31.45)	(75.55)	(31.42)	(75.64)
Bias × Treated		-0.03		-0.04
		(0.05)		(0.05)
Externality × Treated		0.50		0.58
		(1.38)		(1.39)
Cov(bias, treatment effect)		-15,166		-20,555
(standard error)		29,008		29,360
Cov(externality, treatment effect)		160		174
(standard error)		676		677
Var(treatment effect)			36,295	
(standard error)			20,825	
Number of participants	516	516	516	516
Number of observations	1,032	1,032	1,032	1,032

Notes: Table presents estimated ATEs, variances, and covariances for the SmartWay label in the cars experiment. Columns 1 and 2 present fixed coefficient (OLS) versions of equations (12) and (16), respectively, while columns 3 and 4 present the full random coefficient (mixed effects) models. All regressions also include controls for bias and externality as well as indicators for product pairs j and MPL order.

Table A10: **Average Treatment Effects, Variance, and Covariance for Nutrition Facts Label (SSB Experiment)**

	(1) OLS	(2) OLS	(3) Mixed effects	(4) Mixed effects
Treated	-0.45***	-0.52***	-0.45***	-0.52***
	(0.06)	(0.13)	(0.06)	(0.13)
Bias × Treated		0.03		0.03
		(0.05)		(0.05)
Cov(bias, treatment effect)		0.046		0.046
(standard error)		0.078		0.078
Var(treatment effect)			0.796	
(standard error)			0.224	
Number of participants	1,308	1,308	1,308	1,308
Number of observations	3,924	3,924	3,924	3,924

Notes: Table presents estimated ATEs, variances, and covariances for the nutrition facts label in the SSB experiment. Columns 1 and 2 present fixed coefficient (OLS) versions of equations (12) and (16), respectively, while columns 3 and 4 present the full random coefficient (mixed effects) models. All regressions also include controls for bias and externality as well as indicators for product pairs j and MPL order.

Table A11: **Average Treatment Effects, Variance, and Covariance for Stop Sign Warning Label (SSB Experiment)**

	(1) OLS	(2) OLS	(3) Mixed effects	(4) Mixed effects
Treated	-0.34***	-0.53***	-0.34***	-0.53***
	(0.05)	(0.11)	(0.05)	(0.11)
Bias × Treated		0.07*		0.07*
		(0.04)		(0.04)
Cov(bias, treatment effect)		0.108		0.108
(standard error)		0.063		0.063
Var(treatment effect)			0.355	
(standard error)			0.212	
Number of participants	1,327	1,327	1,327	1,327
Number of observations	3,981	3,981	3,981	3,981

Notes: Table presents estimated ATEs, variances, and covariances for the stop sign warning label in the SSB experiment. Columns 1 and 2 present fixed coefficient (OLS) versions of equations (12) and (16), respectively, while columns 3 and 4 present the full random coefficient (mixed effects) models. All regressions also include controls for bias and externality as well as indicators for product pairs j and MPL order.

Table A12: **Average Treatment Effects, Variance, and Covariance for Graphic Warning Label (SSB Experiment)**

	(1) OLS	(2) OLS	(3) Mixed effects	(4) Mixed effects
Treated	-0.51***	-0.90***	-0.51***	-0.90***
	(0.06)	(0.15)	(0.06)	(0.15)
Bias × Treated		0.15***		0.15***
		(0.05)		(0.05)
Cov(bias, treatment effect)		0.234		0.233
(standard error)		0.082		0.082
Var(treatment effect)			1.080	
(standard error)			0.264	
Number of participants	1,318	1,318	1,318	1,318
Number of observations	3,954	3,954	3,954	3,954

Notes: Table presents estimated ATEs, variances, and covariances for the graphic warning label in the SSB experiment. Columns 1 and 2 present fixed coefficient (OLS) versions of equations (12) and (16), respectively, while columns 3 and 4 present the full random coefficient (mixed effects) models. All regressions also include controls for bias and externality as well as indicators for product pairs j and MPL order.

D.2 Alternate Covariance Estimation Strategy for the Drinks Experiment

In this appendix, we consider an alternative strategy to estimating $Cov[\tau, \gamma]$: we estimate $Cov[\tilde{w}, \hat{\gamma}]$, the sample covariance between WTP change and bias. This strategy does not require the the normality assumptions from our primary strategy in Section 3.2.1.

To see when $Cov[\tilde{w}, \hat{\gamma}] = Cov[\tau, \gamma]$, we substitute the model for $\tilde{w}_{ij}$ from equation (12) into $Cov[\tilde{w}, \hat{\gamma}]$:

$$Cov[\tilde{w}_{ij}, \hat{\gamma}_{ij}] = Cov[\tau_{ij} \cdot T_i + \tilde{\epsilon}_{ij}, \hat{\gamma}_{ij}] \tag{74}$$

$$= Cov[\tau_{ij}, \hat{\gamma}_{ij}] + Cov[\tilde{\epsilon}_{ij}, \hat{\gamma}_{ij}]. \tag{75}$$

From this equation, we see that two conditions are sufficient for $Cov[\tilde{w}, \hat{\gamma}] = Cov[\tau, \gamma]$: (i) $Cov[\tau, \hat{\gamma}] = Cov[\tau, \gamma]$ and (ii) $Cov[\tilde{\epsilon}, \hat{\gamma}] = 0$. Condition (i) follows from Assumption 1 in Section 3.2.1, so this strategy is no more restrictive than our primary strategy. In the cars experiment, $\hat{\gamma}_{ij}$ is constructed using baseline WTP w_{ij1}, and is thus mechanically correlated with $\tilde{\epsilon}_{ij}$, violating condition (ii). In the drinks experiment, however, $\hat{\gamma}_{ij}$ is constructed independently of w_{ij}, and thus condition (ii) is plausible.

In the data, $Cov[\tilde{w}, \hat{\gamma}] \approx 0.11$. This is very similar to our primary estimate of 0.13.

E Welfare Analysis Appendix

Table A13: **Parameters and Welfare Analysis: Individual Labels**

(a) **Cars Experiment**

Parameter	(1) Full MPG	(2) Average cost	(3) Personalized cost	(4) SmartWay
$\mathbb{E}[\tau]$	-69	-67	-37	-64
	(32)	(31)	(30)	(31)
$Var[\tau]$	$40,201$	$30,179$	$13,550$	$36,295$
	$(17,339)$	$(17,155)$	$(15,610)$	$(20,825)$
$Cov[\gamma,\tau]$	$12,019$	$-18,247$	$-7,453$	$-20,555$
	$(27,883)$	$(28,034)$	$(28,448)$	$(29,360)$
$Cov[\phi,\tau]$	-165	-19	-184	174
	(691)	(682)	(662)	(677)
$\Delta W(t=0)$	-14.14	6.69	3.42	5.99
$\Delta W(t=t^*)$	-19.04	1.89	0.51	1.33

(b) **Drinks Experiment**

Parameter	(1) Nutrition facts	(2) Stop Sign warning	(3) Graphic warning
$\mathbb{E}[\tau]$	-0.45	-0.34	-0.51
	(0.06)	(0.05)	(0.06)
$Var[\tau]$	0.80	0.35	1.08
	(0.22)	(0.21)	(0.26)
$Cov[\gamma,\tau]$	0.04	0.11	0.24
	(0.08)	(0.06)	(0.08)
$Cov[\phi,\tau]$	0.00	0.00	0.00
$\Delta W(t=0)$	0.12	0.10	0.10
$\Delta W(t=t^*)$	-0.06	-0.04	-0.11

Notes: This table presents parameter estimates and total surplus effects separately for each label in each experiment. Bias $\mathbb{E}[\gamma]$, externality $\mathbb{E}[\phi]$, demand slope D'_p, pass-through ρ, and markup μ are as reported in Table 2. $\Delta W(t=0)$ and $\Delta W(t=t^*)$ are computed using equations (19) and (20), given the parameters reported above. "Unit" is "vehicle-year" for cars and "12-pack" for sugary drinks. Standard errors are in parentheses.