Online Appendix

When Do “Nudges” Increase Welfare?

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

Table of Contents

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 total surplus. These various failures are illustrated in Examples 4, 5, and 6, where the nudge improves deliberative competence but does not increase total 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 total surplus minus the impact on prices. The example in Section 1.2 has shown that even nudges that only "debias" consumers can decrease total 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

$$\begin{aligned}
\frac{dW}{dt} &= -c'(q^*)\frac{dq^*}{dt} && (35)\\
&- \int_{v=p(t)-\gamma-\sigma\tau} f(p(t)-\gamma-\sigma t)(p(t)-\gamma-\sigma\tau)(p'(t)) + \frac{\partial I}{\partial p}\frac{dp}{dt} && (36)\\
&= -c'(q^*)\frac{dq^*}{dt} + D_p p'(t)\left(p(t)-\mathbb{E}_m[\gamma+\sigma\tau]\right) + \rho\frac{\partial I}{\partial p} && (37)\\
&= \frac{dD}{dt}\left(p(t)-\mathbb{E}_m[\gamma+\sigma\tau]\right) - c'(q^*)) + \rho\frac{\partial I}{\partial p}, && (38)
\end{aligned}$$

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

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

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

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

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

or alternatively,

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

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

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

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

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

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

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

Thus

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

Using the multidimensional Leibniz rule, we have that

$$\begin{aligned}
\frac{d}{d\sigma}\int_{v\geq p-\gamma-\sigma\tau} vdF &= -\int_{v=p-\gamma-\sigma\tau}(p-\gamma-\sigma\tau)\left(-\tau+\frac{dp}{d\sigma}\right)dF && (45)\\
&= \mathbb{E}_m[(p-\gamma-\sigma\tau)(-\tau+(1-\rho)\mathbb{E}_m[\tau])]D'_p && (46)\\
&= -p\rho\mathbb{E}_m[\tau]D'_p - \mathbb{E}_m[(\gamma+\sigma\tau)(-\tau+(1-\rho)\mathbb{E}_m[\tau])]D'_p && (47)\\
&= -p\rho\mathbb{E}_m[\tau]D'_p + \rho\mathbb{E}_m[(\gamma+\sigma\tau)\tau]D'_p && (48)\\
&+ (1-\rho)\left(\mathbb{E}_m[(\gamma+\sigma\tau)\tau] - \mathbb{E}_m[\gamma+\sigma\tau]\mathbb{E}_m[\tau]\right)D'_p. && (49)
\end{aligned}$$

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

$$\begin{aligned}
\frac{dW_C}{d\sigma} &= \frac{1}{2}(1-\rho)\frac{\partial}{\partial\sigma}Var_m\left[\gamma+\sigma\tau\right]D'_p + \frac{1}{2}\rho\frac{\partial}{\partial\sigma}\mathbb{E}_m\left[(\gamma+\sigma\tau)^2\right]D'_p && (54)\\
&- \left(\frac{dp}{d\sigma}q^* + p\frac{dq^*}{d\sigma}\right) - p\rho\mathbb{E}_m[\tau]D'_p + \frac{dI}{d\sigma} && (55)\\
&= \frac{1}{2}(1-\rho)\frac{\partial}{\partial\sigma}Var_m\left[\gamma+\sigma\tau\right]D'_p + \frac{1}{2}\rho\frac{\partial}{\partial\sigma}\mathbb{E}_m\left[(\gamma+\sigma\tau)^2\right]D'_p - (1-\rho)\mathbb{E}_m[\tau]q^* + \frac{dI}{d\sigma}. && (56)
\end{aligned}$$

The impact on producer surplus is given by

$$\begin{aligned}
\frac{dW_P}{d\sigma} &= \frac{dp}{d\sigma}q^* + p\frac{dq^*}{d\sigma} - \frac{d}{d\sigma}c(q^*) && (57)\\
&= (1-\rho)\mathbb{E}_m[\tau]q^* - p\rho\mathbb{E}_m[\tau]D'_p - c'(q^*)\rho\mathbb{E}_m[\tau]D'_p && (58)\\
&= (1-\rho)\mathbb{E}_m[\tau]q^* - (p-c'(q^*))\rho\mathbb{E}_m[\tau]D'_p && (59)\\
&= (1-\rho)\mathbb{E}_m[\tau]q^* - \mu\rho\mathbb{E}_m[\tau]D'_p. && (60)
\end{aligned}$$

Putting this together, the impact on total 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 Total 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 Total 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 $\varepsilon_D^j = -\frac{\partial D_j}{dp_j} \cdot \frac{p_j}{D_j}$, and denote the elasticity of demand of product i with respect to the price p_j of product j by ε_D^{ij}. 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 $\theta_j := \frac{p - c'_j(q) - t_j}{p} \varepsilon_D^j$ 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

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

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 total 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**

2020 Subaru Legacy

- 182 horsepower
- 30 miles/gallon
- 2.5 L, 4-cylinder
- 5 seats
- CVT transmission

Details: click here

If gas is free, the maximum I'd pay per year to lease the Subaru Legacy is:

$ ______ per year

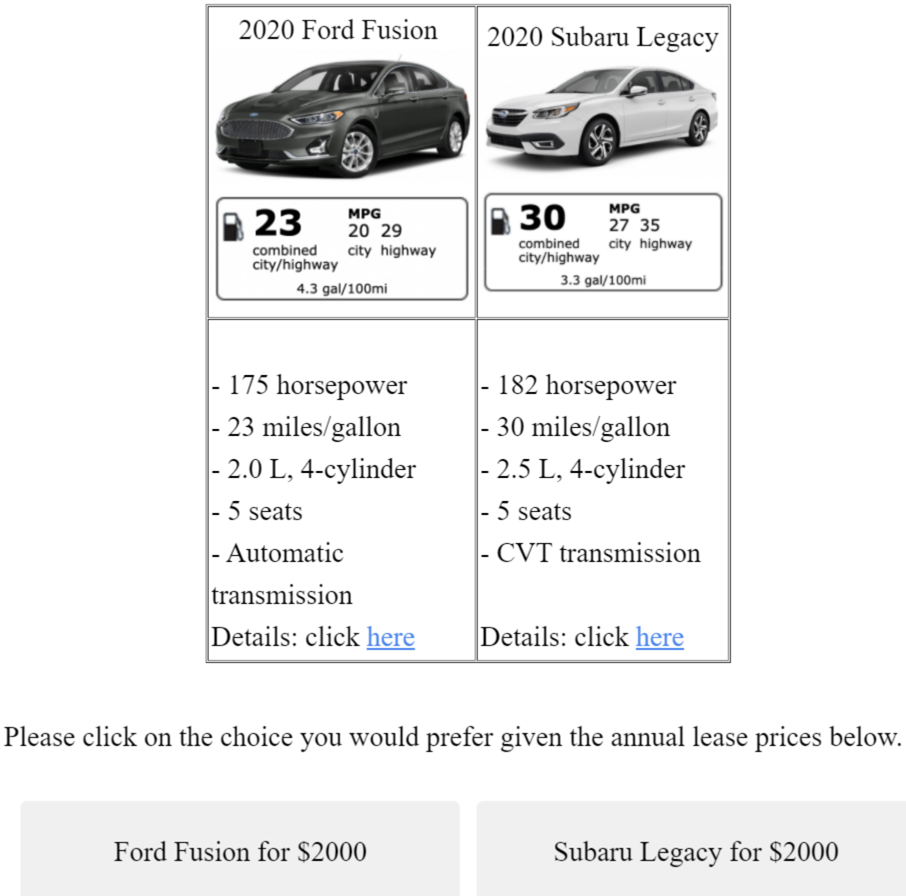
Figure A2: **Cars Experiment: Baseline Multiple Price List**

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**

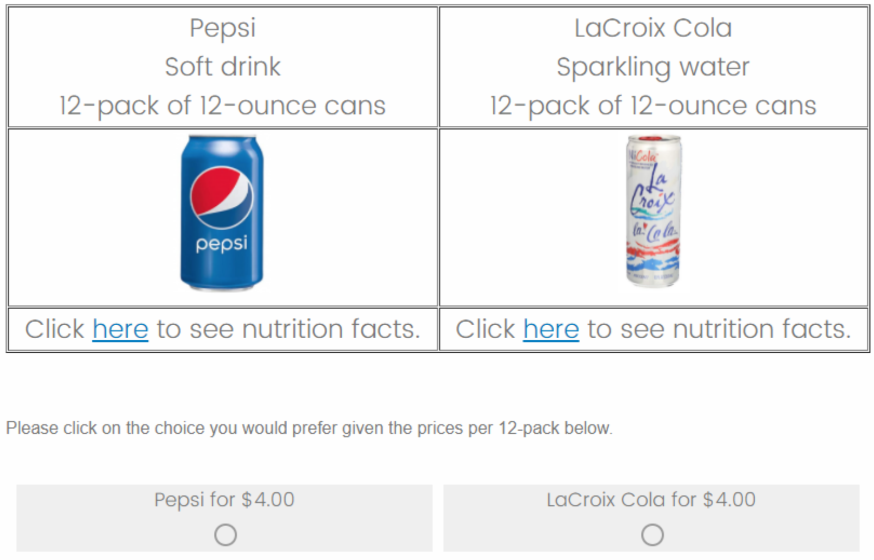
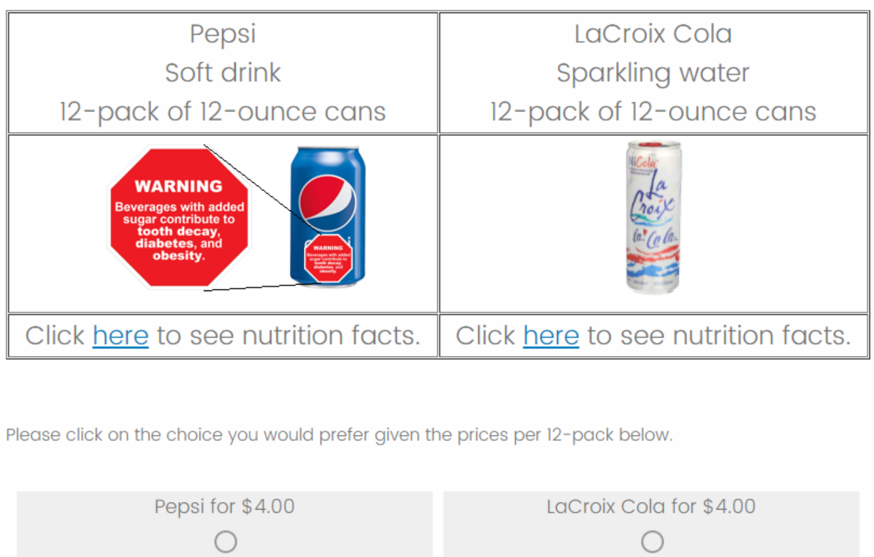


Figure A6: **Drinks Experiment: Endline Multiple Price List with Stop Sign Label**



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 (U.S. Census Bureau 2022). US population average *2019 miles driven* and *2019 gas price* are from the 2017 National Household Travel Survey (Oak Ridge National Laboratory 2018) 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. All demographic data not collected within our survey are from AmeriSpeak's proprietary demographic data panel (National Opinion Research Corporation 2021). 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 (U.S. Census Bureau 2022). *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. All demographic data not collected within our survey are from AmeriSpeak's proprietary demographic data panel (National Opinion Research Corporation 2021).

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: **Cars Experiment: Average Treatment Effects, Variance, and Covariance for Alternative Outlier Approaches**

(a) **Keep All Observations**

	(1) OLS	(2) OLS	(3) Mixed effects	(4) Mixed effects
Treated	-75.71**	-179.72***	-75.48**	-179.66***
	(31.72)	(55.04)	(31.69)	(54.99)
Bias × Treated		0.03		0.03
		(0.03)		(0.03)
Externality × Treated		1.94*		1.94*
		(1.04)		(1.04)
Cov(bias, treatment effect)		2,378,734		2,413,063
(standard error)		(2,664,671)		(2,674,153)
Cov(externality, treatment effect)		11,050		11,062
(standard error)		(4,980)		(4,980)
Var(treatment effect)			49,896	
(standard error)			(43,562)	
Number of participants	2,089	2,089	2,089	2,089
Number of observations	4,178	4,178	4,178	4,178

(b) **Drop Top/Bottom One Percent**

	(1) OLS	(2) OLS	(3) Mixed effects	(4) Mixed effects
Treated	-68.51**	-122.34**	-68.59**	-123.65**
	(28.28)	(52.93)	(28.41)	(52.94)
Bias × Treated		-0.04		-0.05
		(0.05)		(0.05)
Externality × Treated		1.22		1.32
		(0.89)		(0.90)
Cov(bias, treatment effect)		-45,322		-68,674
(standard error)		(79,411)		(79,757)
Cov(externality, treatment effect)		1,366		1,364
(standard error)		(1,150)		(1,151)
Var(treatment effect)			67,262	
(standard error)			(32,211)	
Number of participants	1,792	1,792	1,792	1,792
Number of observations	3,584	3,584	3,584	3,584

Notes: This table presents estimated ATEs, variances, and covariances for the cars 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. In the primary estimates in Table 1, we drop participants in the top or bottom five percent of annual gas cost, WTP if gas is free, estimated bias, or baseline-endline WTP change. Panel (a) instead keeps all observations, and Panel (b) instead drops only participants in the top or bottom one percent.

Table A6: **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

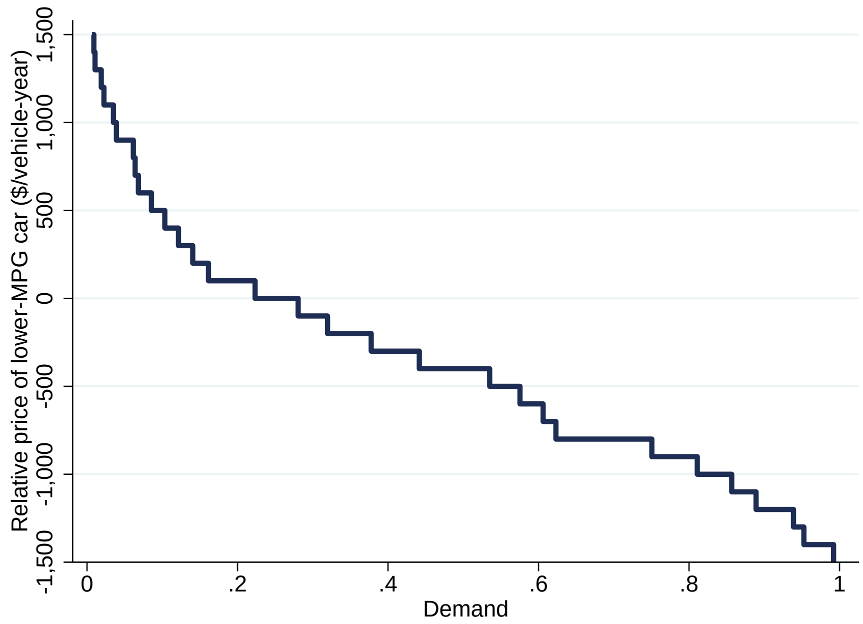
(b) **Drinks Experiment**

	(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

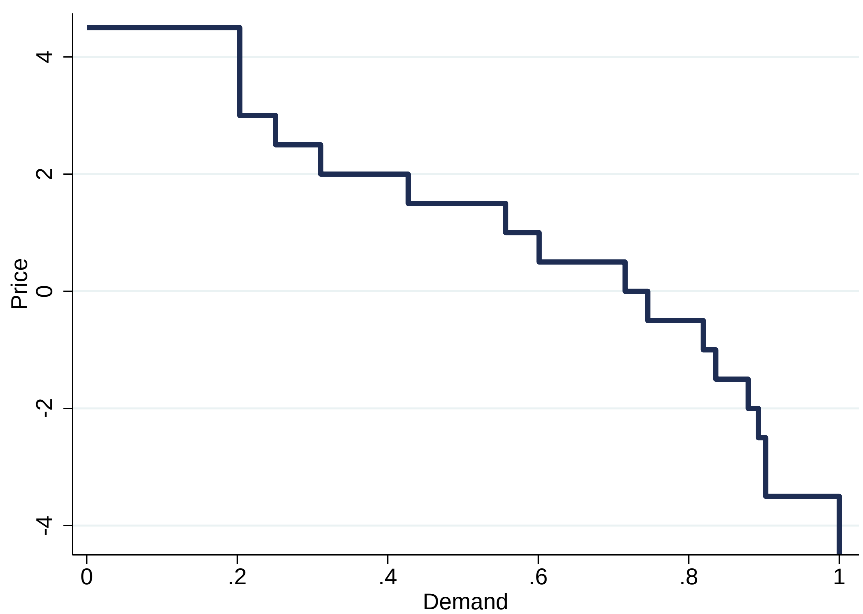
Notes: Panels (a) and (b), respectively, present estimated ATEs, variances, and covariances for the cars experiment and drinks experiment, pooling across all labels. Columns 1 and 2 present fixed coefficient (OLS) versions of equations (12) and (16), respectively, while columns 3 and 4 present the full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



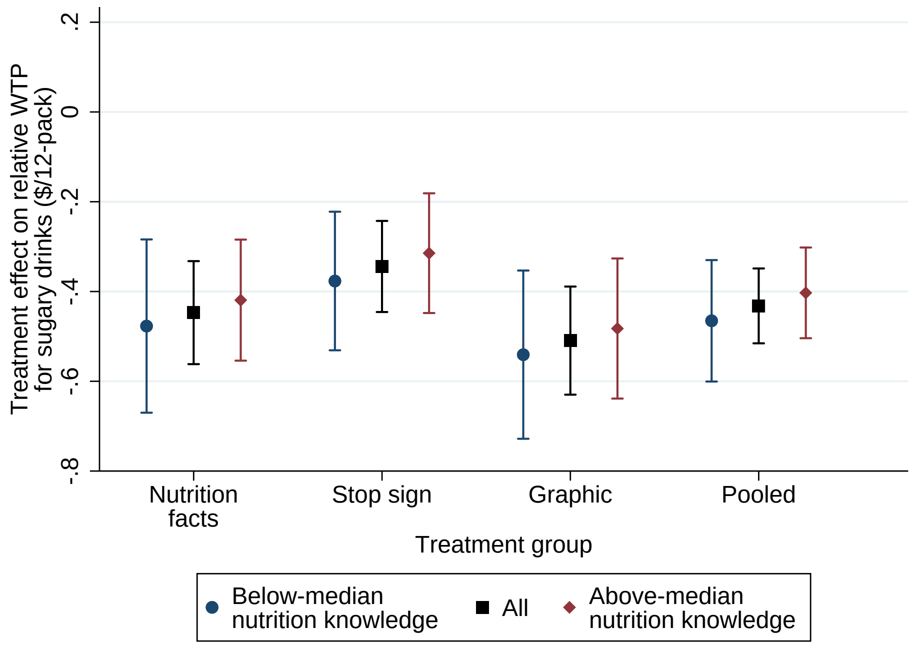
(b) **Drinks Experiment**



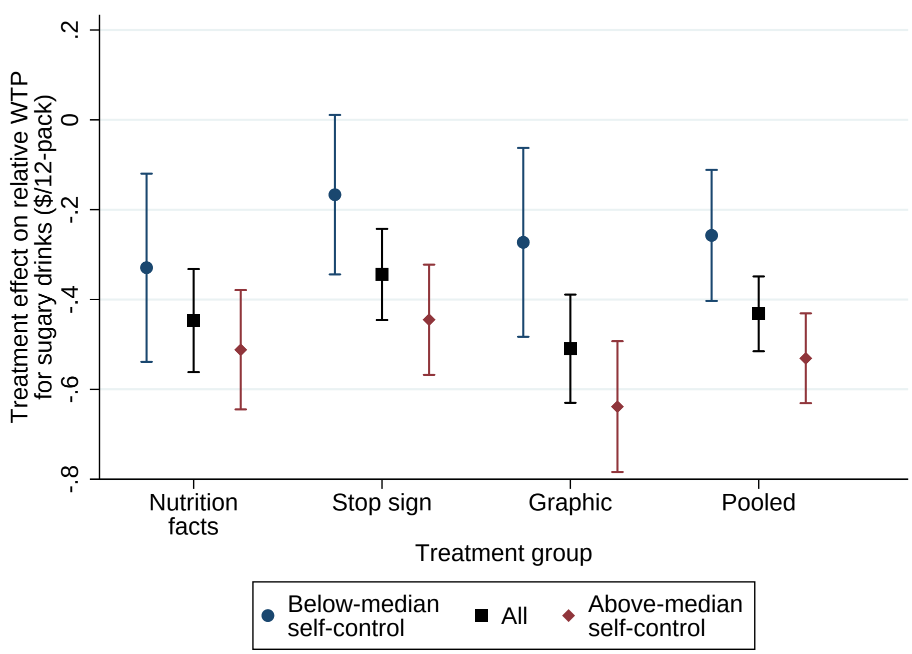
Notes: Panels (a) and (b), respectively, present the baseline demand curves for the cars experiment and drinks experiment.

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

(a) **Nutrition Knowledge**



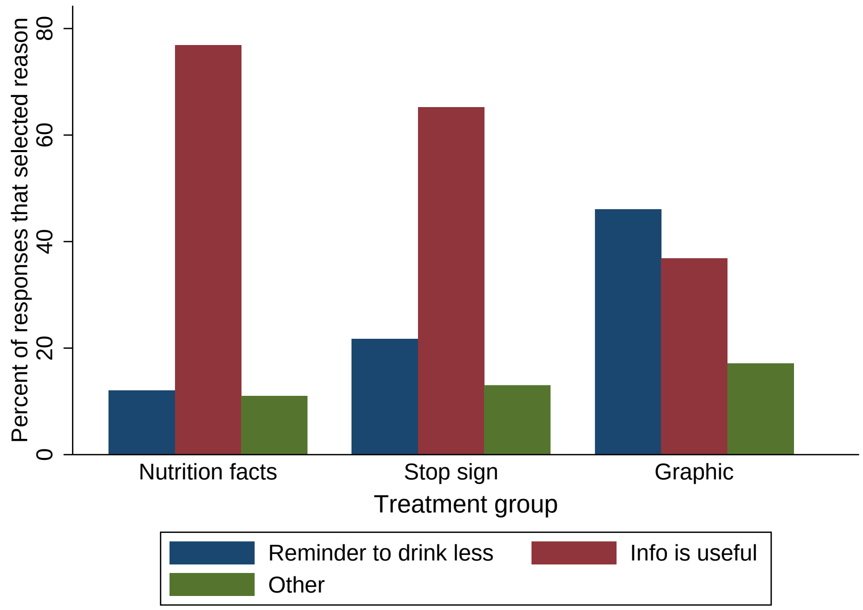
(b) **Self-Control**



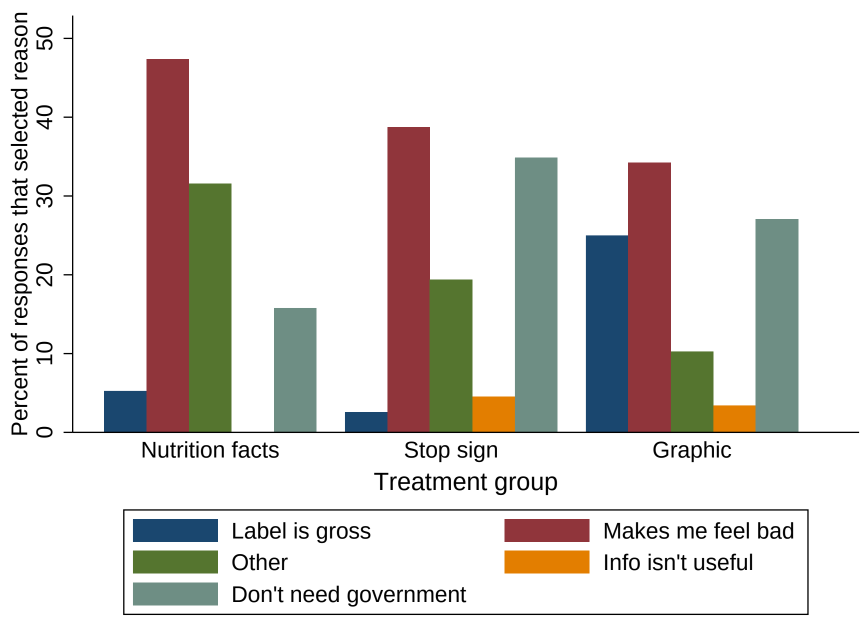
Notes: Panels (a) and (b), respectively, present estimates of equation (12) for the drinks experiment, for subgroups with above- versus below-median nutrition knowledge and self-control. Nutrition knowledge was measured with 28 questions from the General Nutrition Knowledge Questionnaire (GNKQ). Self-control 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." The median response was "mostly," and this is included in the above-median category.

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 A7: **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 A8: **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 A9: **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)			13,549	
(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 A10: **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 A11: **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 A12: **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 A13: **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 A14: **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	13,549	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

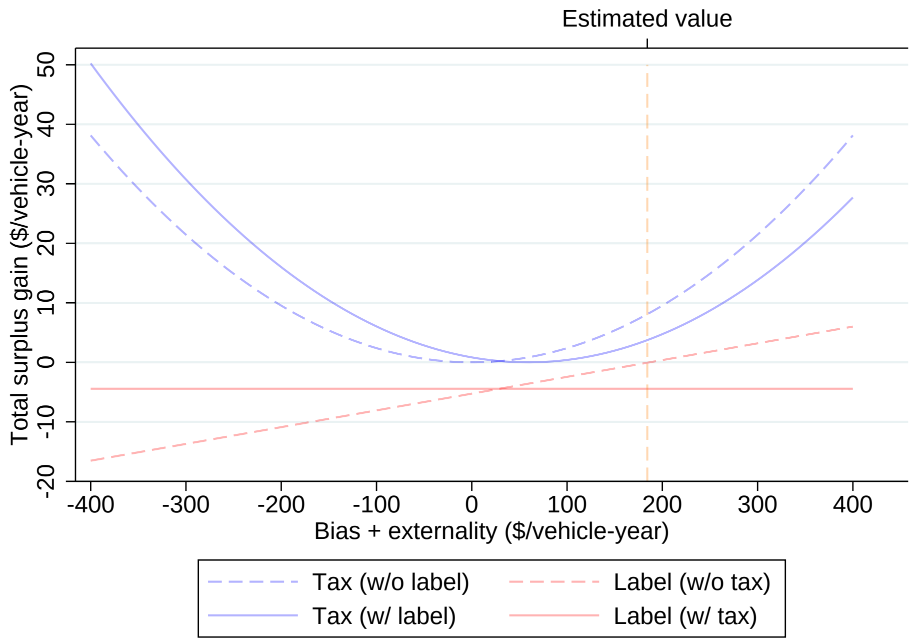
(b) **Drinks Experiment**

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

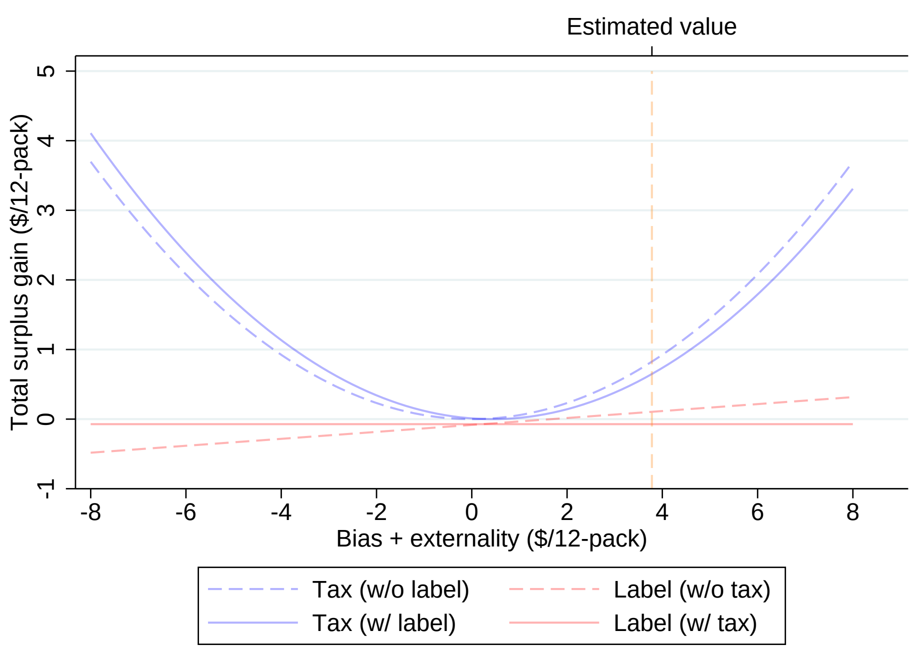
Notes: This table presents parameter estimates and total surplus effects separately for each label in each experiment. Bias $\mathbb{E}[\gamma]$, externality $\mathbb{E}[\phi]$, demand slope D'_p, pass-through ρ, and markup μ are as reported in Table 2. $\Delta W(t=0)$ and $\Delta W(t=t^*)$ are computed 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.

Figure A10: **Total Surplus Under Alternative Bias + Externality Assumptions**

(a) **Cars Experiment**



(b) **Drinks Experiment**



Notes: This figure presents the effects of labels and taxes on total surplus under alternative assumptions for the expected sum of bias plus externalities $\mathbb{E}[\delta]$. 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 $-\frac{1}{2}t^{*2}D'_p$.