When Do “Nudges” Increase Welfare?

“Nudges”

- Examples: information provision, social comparisons, simplification, reminders, framing, defaults, advertising, ...
- Many government “nudge units” (UK, US, DC, Australia, ...), thousands of research papers
- Used to encourage retirement savings, smoking cessation, charitable giving, healthy eating, exercise, social program take-up, organ donation, medication adherence, environmental conservation, ...

How to evaluate nudges?

Popular criterion in practice: Is the average treatment effect (ATE) on behavior in the "right" direction? At low financial cost? (e.g., Benartzi et al. 2017)

Early economic rationale: Offset behavioral bias without distorting decisions of non-biased

- Camerer, Issacharoff, Loewenstein, O'Donoghue, and Rabin (2003): "asymmetric paternalism"
- Thaler and Sunstein (2003): "libertarian paternalism"

How to evaluate nudges?

Popular criterion in practice: Is the average treatment effect (ATE) on behavior in the "right" direction? At low financial cost? (e.g., Benartzi et al. 2017)

Early economic rationale: Offset behavioral bias without distorting decisions of non-biased

- Camerer, Issacharoff, Loewenstein, O'Donoghue, and Rabin (2003): "asymmetric paternalism"
- Thaler and Sunstein (2003): "libertarian paternalism"

All of these are at best loosely related to standard **benefit-cost analysis**

- Emotional and nuisance costs?
- Targeting and efficiency, including relative to taxes

Literature on Nudges

- **General frameworks for welfare analysis**: Farhi and Gabaix (2020), Ambuehl, Bernheim, and Lusardi (2022), Allcott, Cohen, Morrison, Taubinsky (2025), List, Rodemeier, Roy, Sun (2024)
- **Specific welfare evaluations**: Carroll et al. (2009), Handel (2013), Bernheim, Fradkin, Popov (2015), Houde (2018), Damgaard and Gravert (2018), Allcott and Kessler (2019), Allende, Gallego, and Neilson (2019), Thunstrom (2019), Altmann, Grunewald, Radbruch (2022), Barahona, Otero, and Otero (2022), Butera, Metcalfe, Morrison, and Taubinsky (2022) , Choukhmane and Palmer (2023), ...
- **Effects of nudges on behavior**: Madrian and Shea (2001), Gerber and Rogers (2009), Karlan and Gine (2010), Allcott (2011), Bhargava and Manoli (2015), Brewer et al. (2016), Ito, Ida, and Tanaka (2018), Milkman et al. (2021), DellaVigna and Linos (2022), Davis and Metcalf (2016), Allcott and Sweeney (2017), Houde (2018ab), Allcott and Knittel (2019), ... ∞

● Efficient targeting of internalities + externalities ● Emotional, nuisance and bandwidth effects

Today: Targeting internalities + externalities, using the Allcott et al. (2025) framework

Agenda

1. Model and **examples**
2. Experimental implementation and welfare estimates

Model

Model

Consumers:

- Choose whether to buy a good at price p (or which of two goods to buy)
- Good delivers value v. Consumer surplus if buy: $v + (\text{income} - p)$
- Bias γ distorts choice (e.g. inattention, beliefs; or re-interpret as externality)
- Nudge: shifts demand by amount $\sigma\tau$
- Heterogeneous $\{v, \gamma, \tau\}$. Buy if:

$$\underbrace{v}_{\text{value}} - \underbrace{p}_{\text{price}} + \underbrace{\gamma}_{\text{bias}} + \underbrace{\sigma}_{\text{nudge intensity}} \underbrace{\tau}_{\text{nudge effect}} > 0$$

- $D(p)$: aggregate demand. $\varepsilon_D := -pD'(p)/D(p)$: demand elasticity

Model

Consumers:

- Choose whether to buy a good at price p (or which of two goods to buy)
- Good delivers value v. Consumer surplus if buy: $v +$ (income $- p$)
- Bias γ distorts choice (e.g. inattention, beliefs; or re-interpret as externality)
- Nudge: shifts demand by amount $\sigma\tau$
- Heterogeneous $\{v, \gamma, \tau\}$. Buy if:

$$\underbrace{v}_{\text{value}} - \underbrace{p}_{\text{price}} + \underbrace{\gamma}_{\text{bias}} + \underbrace{\sigma}_{\text{nudge intensity}} \underbrace{\tau}_{\text{nudge effect}} > 0$$

- $D(p)$: aggregate demand. $\varepsilon_D := -pD'(p)/D(p)$: demand elasticity

Government: maximizes consumer+producer surplus

- Chooses $\sigma \in \{0, 1\}$, and tax t (paid by producers) redistributed lump-sum

Model

Consumers:

- Choose whether to buy a good at price p (or which of two goods to buy)
- Good delivers value v. Consumer surplus if buy: $v + (\text{income} - p)$
- Bias γ distorts choice (e.g. inattention, beliefs; or re-interpret as externality)
- Nudge: shifts demand by amount $\sigma\tau$
- Heterogeneous $\{v, \gamma, \tau\}$. Buy if:

$$\underbrace{v}_{\text{value}} - \underbrace{p}_{\text{price}} + \underbrace{\gamma}_{\text{bias}} + \underbrace{\sigma}_{\text{nudge intensity}} \underbrace{\tau}_{\text{nudge effect}} > 0$$

- $D(p)$: aggregate demand. $\varepsilon_D := -pD'(p)/D(p)$: demand elasticity

Government: maximizes consumer+producer surplus

- Chooses $\sigma \in \{0, 1\}$, and tax t (paid by producers) redistributed lump-sum

Producers (Weyl and Fabinger 2013 framework)**:**

- Symmetric competition; each firm produces q at cost $c(q)$; constant elasticity-adjusted Lerner index
- $\mu := p - c'(q) - t$: markup. $\rho := \frac{dp}{dt}$: pass-through

Welfare effects of nudge

$\mathbb{E}_m x := \mathbb{E}[x | v + \gamma = p]$: expectation of x over marginal consumers

- Analogous for Var_m and other operators

I. Fixed tax:

$$\Delta W \approx \frac{1}{2}\rho \underbrace{\left((\mathbb{E}_m[\tau + \gamma - \mu - t])^2 - (\mathbb{E}_m[\gamma - \mu - t])^2 \right)}_{\Delta \text{ average distortion squared}} D'_p + \frac{1}{2} \underbrace{(Var_m[\tau + \gamma] - Var_m[\gamma])}_{\Delta \text{ distortion variance}} D'_p$$

II. With optimal tax:

$$\Delta W \approx \frac{1}{2} (Var_m[\tau + \gamma] - Var_m[\gamma]) \cdot D'_p$$

Key statistics

I.Fixed tax:

$$\Delta W \approx \frac{1}{2}\rho \underbrace{\left((\mathbb{E}_m[\tau+\gamma-\mu-t])^2 - (\mathbb{E}_m[\gamma-\mu-t])^2 \right)}_{\Delta \text{ average distortion squared}} D'_p + \frac{1}{2} \underbrace{(Var_m[\tau+\gamma] - Var_m[\gamma])}_{\Delta \text{ distortion variance}} D_p$$

II. With optimal tax:

$$\Delta W \approx \frac{1}{2} (Var_m[\tau+\gamma] - Var_m[\gamma]) \cdot D'_p - \rho\mu\mathbb{E}_m[\tau]D'_p$$

1. $\mathbb{E}_m[\tau]\mathbb{E}_m[\gamma]$
   - Average effect offsetting the bias is generally good
   - But inconsequential in markets with optimal taxes or $\rho = 0$
2. $Cov_m[\tau, \gamma]$
   - Negative covariance between treatment effect and bias is good
3. $Var_m[\tau]$
   - All else equal, noise in τ is bad

Examples

Example 1

- A sugar-sweetened beverage (SSB) market with few distortions: no taxes, no externalities, perfection competition with constant marginal production costs.
- All distortions are psychological:
  - "Oblivious consumers" ignore health harms and health warning labels
    - $\Rightarrow$ Overconsume SSBs before **and** after warning labels introduced
  - "Health nuts" fully aware of health harms (but occasionally have sugary-drinks anyway), but become overly health-obsessed from a health warning label
    - $\Rightarrow$ Optimally consume SSBs pre label, but **under**consume SSBs post label

Example 1

- A sugar-sweetened beverage (SSB) market with few distortions: no taxes, no externalities, perfection competition with constant marginal production costs.
- All distortions are psychological:
  - "Oblivious consumers" ignore health harms and health warning labels
    - $\Rightarrow$ Overconsume SSBs before **and** after warning labels introduced
  - "Health nuts" fully aware of health harms (but occasionally have sugary-drinks anyway), but become overly health-obsessed from a health warning label
    - $\Rightarrow$ Optimally consume SSBs pre label, but **under**consume SSBs post label
- Key characteristics of the market:
  - (i) On average, people overconsume SSBs: $\mathbb{E}[\gamma] > 0$
  - (ii) Labels reduce total consumption of SSBs: $\mathbb{E}[\tau] < 0$

Example 1

- A sugar-sweetened beverage (SSB) market with few distortions: no taxes, no externalities, perfection competition with constant marginal production costs.
- All distortions are psychological:
  - "Oblivious consumers" ignore health harms and health warning labels
    - $\Rightarrow$ Overconsume SSBs before **and** after warning labels introduced
  - "Health nuts" fully aware of health harms (but occasionally have sugary-drinks anyway), but become overly health-obsessed from a health warning label
    - $\Rightarrow$ Optimally consume SSBs pre label, but **under**consume SSBs post label
- Key characteristics of the market:
  - (i) On average, people overconsume SSBs: $\mathbb{E}[\gamma] > 0$
  - (ii) Labels reduce total consumption of SSBs: $\mathbb{E}[\tau] < 0$
- And yet, the labels reduce welfare
  - $Cov[\gamma, \tau] > 0$

Example 2

- Bias for energy-efficient appliances can be positive or negative
  - $2/3$ are “underconsumers” who underestimate energy costs, and the label cuts the underestimation in half
  - $1/3$ are “overconsumers” who overestimate energy costs by the same degree, and the label fully eliminates their overestimation

Example 2

- Bias for energy-efficient appliances can be positive or negative
  - $2/3$ are “underconsumers” who underestimate energy costs, and the label cuts the underestimation in half
  - $1/3$ are “overconsumers” who overestimate energy costs by the same degree, and the label fully eliminates their overestimation
- Key characteristics of the market:
  (i) On average, people under-consume energy-efficient appliances
  (ii) Labels do **not** change total consumption

Example 2

- Bias for energy-efficient appliances can be positive or negative
  - $2/3$ are "underconsumers" who underestimate energy costs, and the label cuts the underestimation in half
  - $1/3$ are "overconsumers" who overestimate energy costs by the same degree, and the label fully eliminates their overestimation
- Key characteristics of the market:
  (i) On average, people under-consume energy-efficient appliances
  (ii) Labels do **not** change total consumption
- And yet, the labels increase welfare
  - $Cov[\tau, \gamma] < 0$

Example 3

- Like examples 1 and 2, but people overestimate the total utility of product by 1 util
- Label drives down this overestimation, on average, but is heterogeneously interpreted
  - Half of the consumers *decrease* their perceived value by 1 util
    - $\Rightarrow$ Correctly value product
  - Half of the consumers *increase* their perceived value by 0.5 utils
    - $\Rightarrow$ Overvalue product by 1.5 utils

Example 3

- Like examples 1 and 2, but people overestimate the total utility of product by 1 util
- Label drives down this overestimation, on average, but is heterogeneously interpreted
  - Half of the consumers *decrease* their perceived value by 1 util
    - $\Rightarrow$ Correctly value product
  - Half of the consumers *increase* their perceived value by 0.5 utils
    - $\Rightarrow$ Overvalue product by 1.5 utils
- Despite reducing overvaluation/over-consumption on average, the label decreases welfare

Example 3

- Like examples 1 and 2, but people overestimate the total utility of product by 1 util
- Label drives down this overestimation, on average, but is heterogeneously interpreted
  - Half of the consumers *decrease* their perceived value by 1 util
    - $\Rightarrow$ Correctly value product
  - Half of the consumers *increase* their perceived value by 0.5 utils
    - $\Rightarrow$ Overvalue product by 1.5 utils
- Despite reducing overvaluation/over-consumption on average, the label decreases welfare
- The nudge creates "noise" in who gets what, making it less likely that consumers with highest v get the product
  - Welfare costs are quadratic in distortion, and $(1.5)^2/2 > 1$
  - This provides intuition for the $Var[\tau]$ statistic

Example 4

- Product supply is fixed (e.g., used car market), so $\rho = 0$
- Consumers are homogeneously biased
- The nudge fully debiases $1/2$ of the consumers, does not affect the other $1/2$

Example 4

- Product supply is fixed (e.g., used car market), so $\rho = 0$
- Consumers are homogeneously biased
- The nudge fully debiases $1/2$ of the consumers, does not affect the other $1/2$
- Despite unambiguously helping consumers choose better, the nudge is bad for welfare

Example 4

- Product supply is fixed (e.g., used car market), so $\rho = 0$
- Consumers are homogeneously biased
- The nudge fully debiases $1/2$ of the consumers, does not affect the other $1/2$
- Despite unambiguously helping consumers choose better, the nudge is bad for welfare
- With $\rho = 0$, all that matters is the impact on distortion variance: $Var[\gamma + \tau] - Var[\gamma]$
- Intuition:
  - Markets with $\rho = 0$ (or $\rho < 1$, more generally) are partially self-correcting: demand shocks from bias pass through to prices
  - Heterogeneity in bias creates misallocation, but the average bias is irrelevant

Example 5

- As in Ex 4, consumers homogeneously biased and nudge debiases $1/2$ of the consumers
  - For illustration: consumers initially overvalue the product by an amount equal to \$1 per unit
- Government can impose a tax to counteract the bias

Example 5

- As in Ex 4, consumers homogeneously biased and nudge debiases $1/2$ of the consumers
  - For illustration: consumers initially overvalue the product by an amount equal to \$1 per unit
- Government can impose a tax to counteract the bias
- Welfare is maximized by setting an optimal tax $t = \$1$ and *not* using the nudge

Example 6

- Due to market power, prices are $1 above marginal costs
- Consumers overestimate utility from product by $1

Example 6

- Due to market power, prices are $1 above marginal costs
- Consumers overestimate utility from product by $1
- This bias is good for welfare!

Key statistics

I.Fixed tax:

$$\Delta W \approx \frac{1}{2}\rho \underbrace{\left((\mathbb{E}_m[\tau+\gamma-\mu-t])^2 - (\mathbb{E}_m[\gamma-\mu-t])^2 \right)}_{\Delta \text{ average distortion squared}} D'_p + \frac{1}{2} \underbrace{(Var_m[\tau+\gamma] - Var_m[\gamma])}_{\Delta \text{ distortion variance}} D_p$$

II. With optimal tax:

$$\Delta W \approx \frac{1}{2} (Var_m[\tau+\gamma] - Var_m[\gamma]) \cdot D'_p - \rho\mu\mathbb{E}_m[\tau]D'_p$$

1. $\mathbb{E}_m[\tau]\mathbb{E}_m[\gamma]$
   - Average effect offsetting the bias is generally good
   - But inconsequential in markets with optimal taxes or $\rho = 0$
2. $Cov_m[\tau, \gamma]$
   - Negative covariance between treatment effect and bias is good
3. $Var_m[\tau]$
   - All else equal, noise in τ is bad

Experimental designs and data

Experimental designs

Two sets of common product labels, simplified to binary choices:

- Fuel economy labels: low- vs. high-MPG cars
- Health labels: sugary vs. zero-calorie drinks

Experimental designs

Two sets of common product labels, simplified to binary choices:

- Fuel economy labels: low- vs. high-MPG cars
- Health labels: sugary vs. zero-calorie drinks

Online incentivized experiments:

1. introductory questions to measure bias
2. baseline multiple price list (MPL)
3. endline MPL with randomized information labels

Drinks experiment

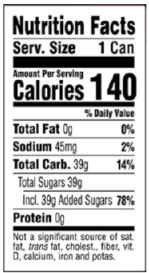
Drinks experiment overview

- Sample: 2,619 people recruited from Facebook from October–December 2021
- Binary choices:
  - Sugary drinks: Minute Maid Lemonade, Coke, Pepsi, Seagrams Ginger Ale, Sprite, Crush
  - Sugar-free drinks: LaCroix, Bubly, 365 sparkling waters with similar flavor

Drinks experiment overview

- Sample: 2,619 people recruited from Facebook from October–December 2021
- Binary choices:
  - Sugary drinks: Minute Maid Lemonade, Coke, Pepsi, Seagrams Ginger Ale, Sprite, Crush
  - Sugar-free drinks: LaCroix, Bubly, 365 sparkling waters with similar flavor
- Incentive compatible: randomly selected 22% of participants, sent them the drinks they chose on a random MPL question

Labels



Nutrition facts label

Stop sign warning label

Graphic warning label

Baseline MPL

Pepsi Soft drink 12-pack of 12-ounce cans	LaCroix Cola Sparkling water 12-pack of 12-ounce cans
	
Click here to see nutrition facts.	Click here to see nutrition facts.

Please click on the choice you would prefer given the prices per 12-pack below.

Pepsi for $4.00	LaCroix Cola for $4.00
○	○

Endline MPL with stoplight label

Pepsi Soft drink 12-pack of 12-ounce cans	LaCroix Cola Sparkling water 12-pack of 12-ounce cans
	
Click here to see nutrition facts.	Click here to see nutrition facts.

Please click on the choice you would prefer given the prices per 12-pack below.

Pepsi for $4.00	LaCroix Cola for $4.00
○	○

Bias measurement

- Possible biases: lack of nutrition knowledge or self-control

- Baseline survey (3 days before MPLs): elicit bias proxies
  - *Nutrition knowledge*: score on 28-question General Nutrition Knowledge Questionnaire (Kliemann et al. 2016)
  - *Self-control*: agreement with “I drink soda pop or other sugar-sweetened beverages more often than I should”

Bias measurement

- Possible biases: lack of nutrition knowledge or self-control
- Baseline survey (3 days before MPLs): elicit bias proxies
  - *Nutrition knowledge*: score on 28-question General Nutrition Knowledge Questionnaire (Kliemann et al. 2016)
  - *Self-control*: agreement with "I drink soda pop or other sugar-sweetened beverages more often than I should"
- Bias estimate from Allcott, Lockwood, and Taubinsky (QJE 2019):
  - Step 1: Calculate predicted consumption if have perfect self-control and nutrition knowledge (controlling for tastes, demographics, etc)
  - Step 2: Translate consumption wedge to $$ by estimating the price elasticity of demand

$$\widehat{\text{Bias}}_i = \kappa_1 \cdot (\text{nutrition knowledge})_i + \kappa_2 \cdot (\text{self-control})_i + \kappa_0$$

Intuition for identifying treatment effect averages and heterogeneity

- $\mathbb{E}[\tau]$: Average ΔWTP in treated versus control groups
- $Cov[\tau, \gamma]$: Covariance between ΔWTP and bias estimates $\hat{\gamma}_i$ in treated versus control groups
- $Var[\tau]$: How much more variance is there in ΔWTP in treated versus control groups

Formal regression model

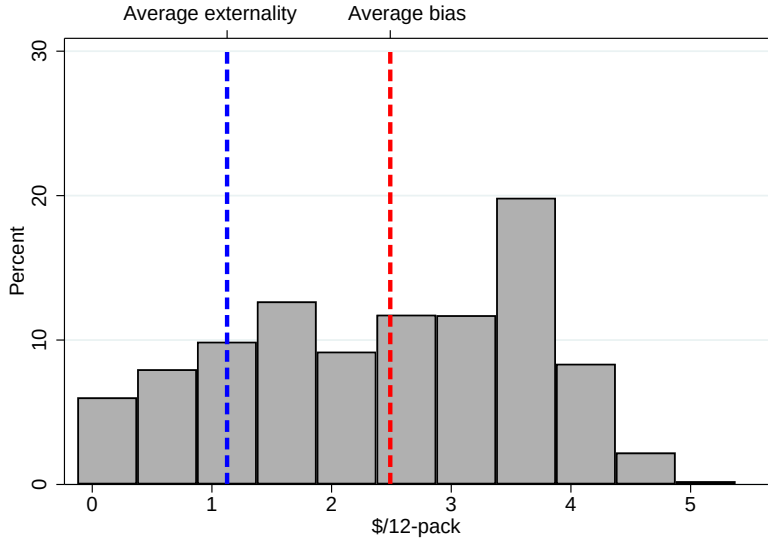
Standard mixed effects regression model:

$$w_{ij1} - w_{ij0} = \eta_{ij} T_i + \alpha_1 \hat{\gamma}_{ij} T_i + \beta_1 \hat{\gamma}_{ij} + \beta_{0i} + \nu_{ij}$$

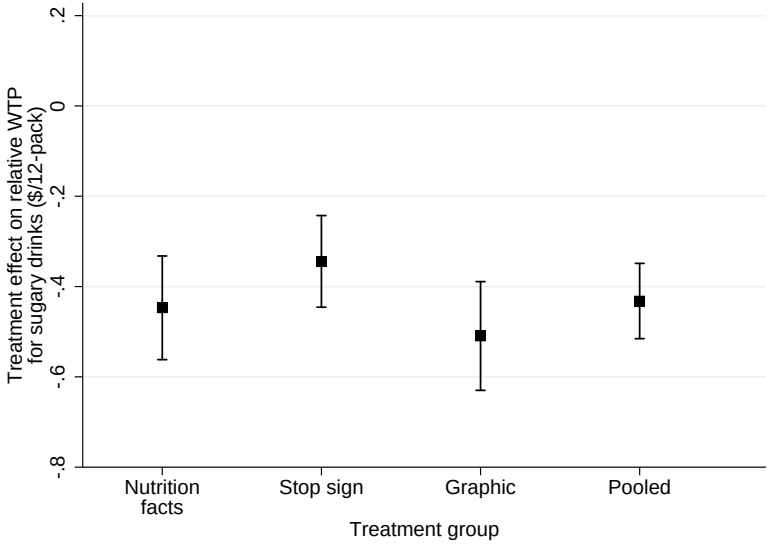
- Random coefficients: η_{ij}, β_{0i}
  - Mean of η_{ij} is average treatment effect
  - Variance of η_{ij} is variance of treatment effects
  - Randomness in β_{0i} allows for individual-level variance in mean reversion
- Fixed coefficients: α_1, β_1
  - α_1 identifies $Cov[\hat{\gamma}, \tau] = Cov[\gamma, \tau]$
  - β_1 controls for any relationship between mean reversion and bias

Experimental results

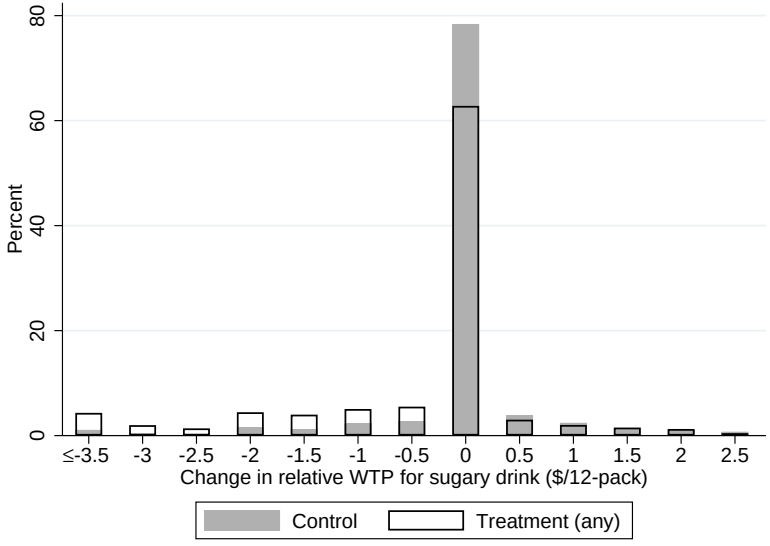
Average externality and distribution of estimated bias



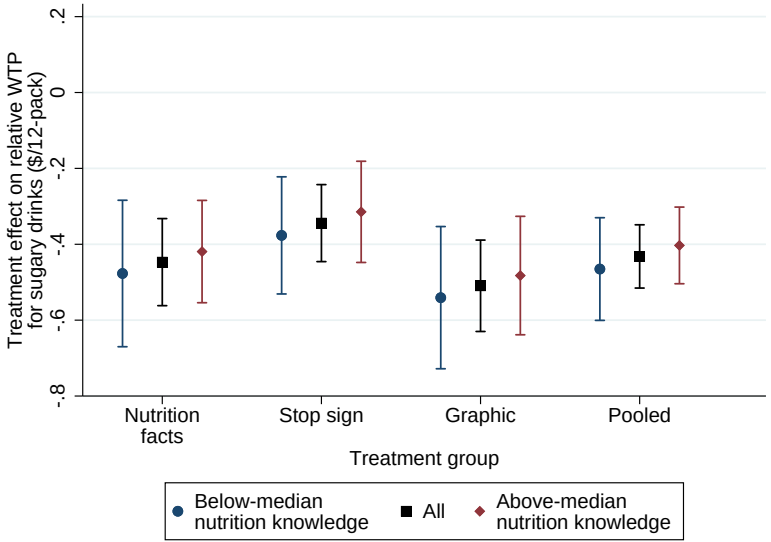
Average effects



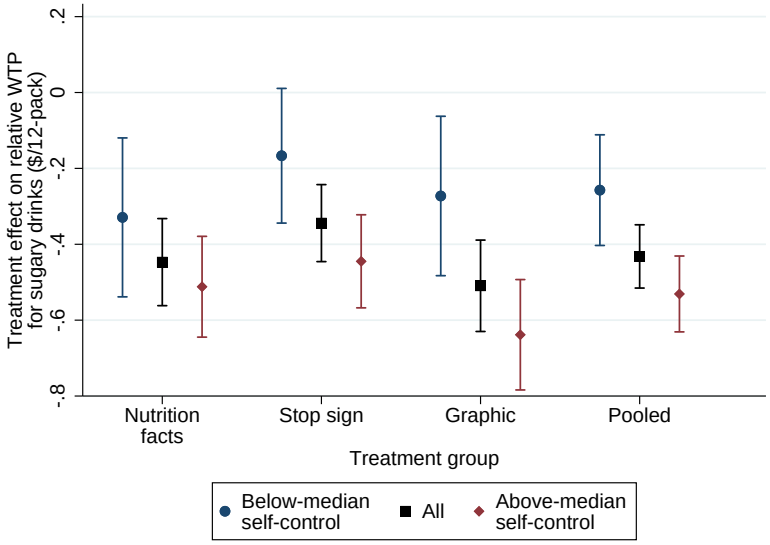
Heterogeneous WTP changes in treatment and control



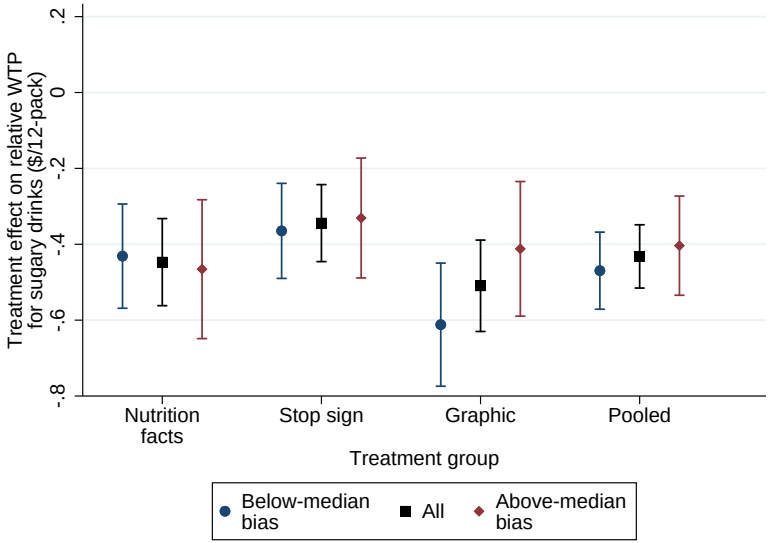
Targeting: effects for above- and below-median nutrition knowledge



Targeting: effects for above- and below-median self-control



Targeting: effects for above- and below-median bias

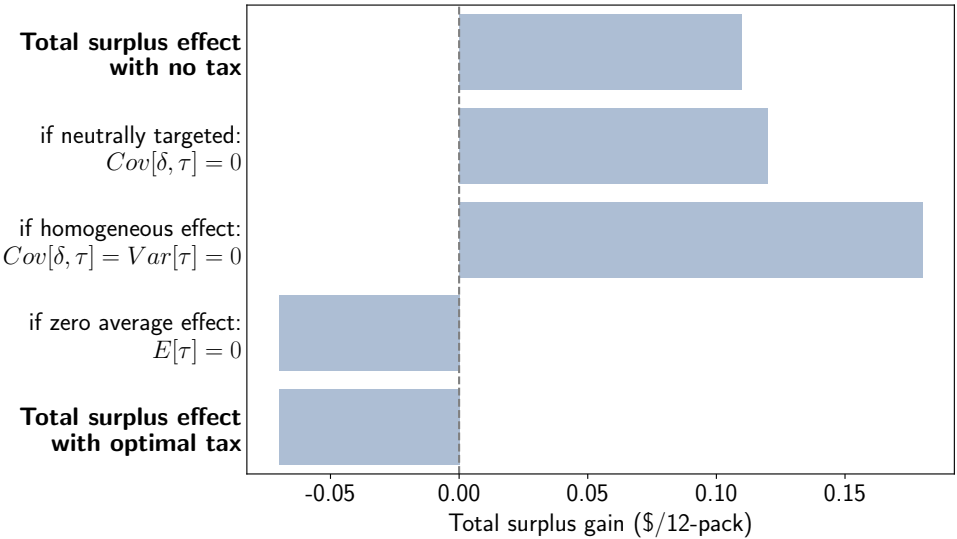


Welfare analysis

Parameter estimates

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

Welfare effects of sugary drink labels (pooled)



Welfare effects of labels (pooled)

Parameter	Description	(1) Cars experiment	(2) Drinks experiment
$\Delta W(t=0)$	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

Conclusion

Recap and takeaways

Recap:

- Embed nudges in standard public finance framework
- Evaluate canonical information labels using randomized experiments

Takeaways:

- Nudges can decrease welfare (by adding variance/noise) even if they change behavior “in the right direction”
- ATEs are not sufficient statistics for welfare
  - And are nearly irrelevant w/ well-set taxes or low pass-through
- And Libertarian or Asymmetric paternalism diverge from standard consumer surplus metrics as well

Implications for design

In addition to moving average behavior in the “right” direction, use nudges that:

- Are well-targeted: have larger effects on more biased people
- Not noisy: Similarly interpreted by people with similar biases

Food for thought: which existing nudges do that?