

# Identifying the Impact of Hypothetical Stakes on Experimental Outcomes and Treatment Effects

Jack Fitzgerald

Vrije Universiteit Amsterdam

November 19, 2024



# The Way It Is

Experimental economics has a long-standing norm of incentivizing experimental choices with real stakes (see e.g., Bardsley et al. 2009)

- ▶ We do this because theory shows that financial incentives can drown out known experimental biases and ensure generalizability (cf. Smith 1976; Smith 1982)

Incentivization is a key mechanism by which experimental economics differentiates itself from other experimental social sciences (Camerer & Hogarth 1999; Hertwig & Ortmann 2001)

# Oh, the Times They Are A-Changin'

These norms have recently started to shift

- ▶ Large online/population surveys elicit economic preferences w/ hypothetical-stakes experiments (e.g., Global Preferences Survey; see Falk et al. 2018)
- ▶ Top economics journals are increasingly publishing results from hypothetical-stakes experiments (Golsteyn, Grönqvist, & Lindahl 2014; Cadena & Keys 2015; Kuziemko et al. 2015; Alesina, Stantcheva, & Teso 2018; Sunde et al. 2022; Stango & Zinman 2023)

# The Evidence Behind the Trend

This trend is bolstered by recent evidence showing that some experimental outcomes do not stat. sig. differ between real-stakes and hypothetical-stakes conditions (Brañas-Garza, Kujal, & Lenkei 2019; Matousek, Havranek, & Irsova 2022; Alfonso et al. 2023; Enke et al. 2023; Hackethal et al. 2023)

- ▶ These studies are often being directly used to justify utilizing results from hypothetical-stakes experiments (e.g., Brañas-Garza et al. 2021; Brañas-Garza et al. 2023)

This evidence is affecting thinking at the highest levels of experimental economics

- ▶ *Experimental Economics* has an upcoming special issue on incentivization; citing some of this evidence, the editorial board states in the announcement:

*“There is good rationale for incentivized experiments, but recently there has been evidence that incentivization may not always matter.”*

# This Project

Econometrically, this recent literature does not rule out the hypothetical biases we care about most

- ▶ Classical hypothetical bias studies do not identify hypothetical biases on treatment effects (TEs) or their standard errors (SEs)
- ▶ **Intuition:** TE-relevant hypothetical biases are interaction effects, but most hypothetical bias studies only estimate TE-irrelevant differences in means

I show empirically that inferring conclusions about TE-relevant biases from estimates of TE-irrelevant biases can be quite misleading

- ▶ In re-analyses of three recent hypothetical bias experiments, TE-irrelevant biases often yield completely different conclusions than TE-relevant biases
- ▶ TE-irrelevant biases can even hold the opposite sign of TE-relevant biases

Hypothetical bias experiments which try to 'pave the way' for researchers to use hypothetical-stakes experiments are thus unproductive, uninformative, and potentially misleading

- ▶ Norms in favor of (probabilistically) real stakes are still a good idea

# A Simple Taxonomy of Experiments

I distinguish between two key types of experiments

## Elicitation Experiment

An experimental procedure is used to elicit an outcome  $Y$ . The statistic of interest is a descriptive statistic (usually a mean) of  $Y$ , rather than any TE.

## Intervention Experiment

An intervention  $D$  is introduced. The statistic of interest is the TE of  $D$  on  $Y$ .

# The Taxonomy Applied (1/2)

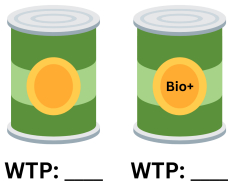


**WTP:** \_\_\_\_

Suppose we're interested in finding the average customer's willingness to pay (WTP) for a product

- ▶ **Experiment:** Run Becker-DeGroot-Marschak (BDM; 1964) procedure to obtain average WTP
- ▶ Clearly an experiment, but no TE at play – we're just interested in descriptive statistics on WTP
- ▶ This is an elicitation experiment

## The Taxonomy Applied (2/2)



Now suppose we're interested in finding the average effect of some product feature on WTP

- ▶ **Experiment:** Randomize product characteristic between/within subjects and run BDM to obtain WTPs; difference in average WTPs is the TE of interest
- ▶ This is an intervention experiment, even if we're still interested in average WTP within the two treatment conditions



## Some More Vocab

### Stakes Condition

$S_i$  indicates that subject  $i$  faces real stakes with probability  $p'$  instead of probability  $p$ . I.e., for  $p, p' \in [0, 1]$ ,

$$S_i = \begin{cases} 0 & \text{if subject } i\text{'s stakes are real with probability } p \\ 1 & \text{if subject } i\text{'s stakes are real with probability } p' \end{cases}$$

Ordinarily,  $p = 1$  and  $p' = 0$ , so  $S_i = 1$  ( $S_i = 0$ ) indicates pure hypothetical (real) stakes.

### Hypothetical Bias

The effect of stakes condition  $S_i$  on our statistic of interest.

# What Did Early Experimental Economists Care About?

Pre-1960, experimental economics was dominated by elicitation experiments

- ▶ This is because early economic experiments were aimed at testing prevailing theories on utility, equilibria, and rationality (see Thurstone 1931; Chamberlin 1948; Rousseas & Hart 1951; Allais 1953; Mosteller 1953; Flood 1958)

This history is important because it influenced the biases that experimental economists cared about when norms on experimental stakes first developed

- ▶ The influential Wallis & Friedman (1942) critique inspired leading experimental economists to incentivize experiments (Ortmann 2016; Svorencik & Maas 2016)
- ▶ Norms on incentivizing experiments with real stakes solidified by the late 1950s (Roth 1995)

Two rationales for real experimental stakes emerged from this early literature

## Bias on Estimates: Classical Hypothetical Bias

Hypothetical stakes may bias the average preference/behavior elicited from a sample

- ▶ This implies that hypothetical stakes bias  $\mathbb{E}[Y_i]$

### Classical Hypothetical Bias (CHB)

Let  $Y_i(S)$  be participant  $i$ 's potential outcome of  $Y_i$ , dependent on  $S_i \in \{0, 1\}$ . Then

$$\text{CHB} = \mathbb{E}[Y_i(1) - Y_i(0)].$$

I.e., CHB is the average marginal effect of  $S_i$  on  $Y_i$ .

Easy to verify that if  $S_i$  is randomized/unconfounded, CHB is just  $\delta$  in the following regression:

$$Y_i = \alpha + \delta S_i + \epsilon_i$$

# Documented CHB in Elicitation Experiments

CHB has been widely documented in elicitation experiments

- ▶ **Systematic evidence:** Real stakes impact average behavior/performance in a substantial proportion of hypothetical bias experiments (Smith & Walker 1993; Camerer & Hogarth 1999; Hertwig & Ortmann 2001; Harrison & Rutström 2008)
- ▶ **Popular economic experiments:** Ultimatum games (Sefton 1992), public goods games (Cummings et al. 1997), auctions (List 2001), multiple price lists (Harrison et al. 2005), contingent valuation (Murphy et al. 2005; Hausman 2012)

The recent hypothetical bias literature is largely providing evidence on CHB (e.g., Brañas-Garza, Kujal, & Lenkei 2019; Brañas-Garza et al. 2021; Matousek, Havranek, & Irsova 2022; Brañas-Garza et al. 2023; Hackethal et al. 2023)

## Bias on Noise: Outcome Standard Deviation Bias

Real stakes may reduce noise by increasing attention and effort

- ▶ This implies that hypothetical stakes bias the standard deviation (SD) of  $Y_i$

### Outcome Standard Deviation Bias (OSDB)

Let  $\sigma_{Y_i}(S)$  be the standard deviation of  $Y_i$  given stakes condition  $S \in \{0, 1\}$ . Then

$$\text{OSDB} = \mathbb{E}[\sigma_{Y_i}(1) - \sigma_{Y_i}(0)].$$

In a hypothetical bias experiment, we can get an OSDB point estimate by taking the difference in  $\sigma_{Y_i}$  between stakes conditions, and a standard error via bootstrap

- ▶ There is systematic evidence that hypothetical stakes increase outcome noise (Smith & Walker 1993; Camerer & Hogarth 1999; Hertwig & Ortmann 2001)

# How Much Do We Care About These Hypothetical Biases?

CHB and OSDB are fully-informative hypothetical bias measures in elicitation experiments

- ▶ Because most early economic experiments were elicitation experiments, CHB and OSDB were the only hypothetical biases that mattered to most early experimental economists
- ▶ This early focus on CHB and OSDB has carried over into modern measurement of hypothetical bias (see Brañas-Garza, Kujal, & Lenkei 2019; Brañas-Garza et al. 2021; Matousek, Havranek, & Irsova 2022; Brañas-Garza et al. 2023; Hackethal et al. 2023)

**However, these bias measures are completely irrelevant for intervention experiments**

- ▶ We can model CHB and ODSB while completely ignoring any intervention  $D_i$
- ▶ Understanding hypothetical bias on *treatment effects* requires bias measures that incorporate  $D_i$  (Guala 2001)

## A Better Statistical Framework

I consider a simple linear heterogeneous treatment effects framework (Guala 2001):

$$Y_i = \alpha + \beta_1 D_i + \beta_2 S_i + \beta_3 (D_i \times S_i) + \mu_i$$

Suppose that we randomize both  $D_i$  and  $S_i$  in a factorial experiment

- ▶ This renders both unconfounded:  $\mu_i \perp \{D_i, S_i\}$

Then if  $Y_i(D, S)$  is the potential outcome of  $Y_i$  given intervention dummy  $D_i \in \{0, 1\}$  and stakes condition  $S_i \in \{0, 1\}$ , we can model participant  $i$ 's treatment effect as

$$\tau_i = Y_i(1, S) - Y_i(0, S) = \begin{cases} \beta_1 & \text{if } S_i = 0 \\ \beta_1 + \beta_3 & \text{if } S_i = 1 \end{cases}$$

# Bias on TE Estimates: Interactive Hypothetical Bias

## Interactive Hypothetical Bias (IHB)

Let  $\tau_i(S)$  be the TE of  $D_i$  on  $Y_i$  given stakes condition  $S_i \in \{0, 1\}$ . Then

$$\text{IHB} = \mathbb{E} [\tau_i(1) - \tau_i(0)].$$

Consider the previous data-generating process:

$$Y_i = \alpha + \beta_1 D_i + \beta_2 S_i + \beta_3 (D_i \times S_i),$$

Here IHB is just the interaction effect between the intervention and the stakes condition:

$$\mathbb{E} [\tau_i(1) - \tau_i(0)] = \underbrace{\mathbb{E} [Y_i(1, 1) - Y_i(0, 1)]}_{\beta_1 + \beta_3} - \underbrace{\mathbb{E} [Y_i(1, 0) - Y_i(0, 0)]}_{\beta_1} = \beta_3$$



## What Does CHB Identify?

Recall the data-generating process:

$$Y_i = \alpha + \beta_1 D_i + \beta_2 S_i + \beta_3 (D_i \times S_i)$$

We can write the *individual* marginal effect of  $S_i$  on  $Y_i$  as

$$\delta_i = Y_i(D, 1) - Y_i(D, 0) = \begin{cases} \beta_2 & \text{if } D_i = 0 \\ \beta_2 + \beta_3 & \text{if } D_i = 1 \end{cases}$$

By definition, CHB is the *average* marginal effect of  $S_i$  on  $Y_i$ , which we can get via expectation:

$$\mathbb{E}[\delta_i] = \beta_2 + \mathbb{E}[D_i]\beta_3$$

# CHB Does Not Identify IHB

$$\mathbb{E}[\delta_i] = \beta_2 + \mathbb{E}[D_i]\beta_3$$

**IHB** cannot be identified from the **CHB** estimated in traditional hypothetical bias experiments

- ▶ Identifying **IHB** from **CHB** requires knowing at least two of  $\beta_2$ ,  $\mathbb{E}[D_i]$ , and  $\beta_3$
- ▶ However, both  $\mathbb{E}[D_i]$  and  $\beta_3$  are undefined in experiments where no  $D_i$  is varied
- ▶ You need a factorial experiment that varies both  $S_i$  and  $D_i$  to identify  $\beta_3$

Inferring **IHB** from **CHB** without a factorial experiment can yield misleading conclusions

- ▶ If  $\|\beta_2\|$  is large but  $\beta_3 = 0$ , then **CHB** is large while **IHB** is zero
- ▶ If  $\mathbb{E}[\delta_i] = 0$  but  $\|\beta_3\|$  is large, then **CHB** is zero while **IHB** is large
- ▶ If  $\text{sgn}(\beta_2) \neq \text{sgn}(\beta_3)$ , then **CHB** and **IHB** can hold opposite signs
- ▶ **IHB**  $\neq$  **CHB** whenever  $\beta_3 \neq \frac{\beta_2}{1 - \mathbb{E}[\delta_i]}$ , which is virtually always true

# What About Meta-Analyses?

Some meta-analyses take the following approach (e.g., Li, Maniadis, & Sedikides 2021)

- ▶ Look at a popular TE that's been studied both with and without real stakes
- ▶ Gather standardized TE estimates under real stakes  $\tau(p)$  and hypothetical stakes  $\tau(p')$
- ▶ Compute IHB by calculating  $\mathbb{E}[\tau(p')] - \mathbb{E}[\tau(p)]$

**This meta-analysis does not generally provide clean evidence on IHB**

- ▶ To identify IHB, we need joint unconfoundedness over both  $D_i$  and  $S_i$  ( $D_i, S_i \perp \mu_i$ )
- ▶ Easy assumption *within* factorial experiments via randomization
- ▶ However, *across* experiments,  $S_i$  is generally endogenously assigned
- ▶ E.g., economics labs incentivize more often than psychology labs, and psychology students respond significantly differently to the same treatments than do economics students (Van Lange, Schippers, & Balliet 2011; van Anandel, Tybur, & Van Lange 2016)

## Bias on TE Noise: Treatment Effect Standard Error Bias

### Treatment Effect Standard Error Bias (TESEB)

Let  $\tau_i(S)$  be the TE of  $D_i$  on  $Y_i$  given stakes condition  $S_i \in \{0, 1\}$ . Then

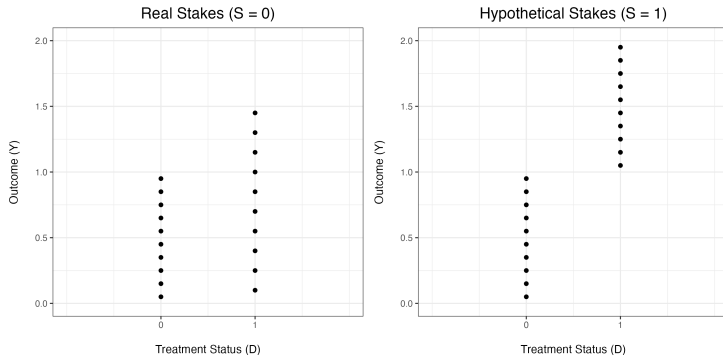
$$\text{TESEB} = \mathbb{E} [\text{SE}(\tau(1)) - \text{SE}(\tau(0))].$$

In a factorial hypothetical bias experiment that varies both  $D_i$  and  $S_i$ , we can estimate TESEB by taking differences in  $\text{SE}(\tau)$  between stakes conditions

- ▶ We can get SEs via bootstrapping

**OSDB does not identify TESEB**

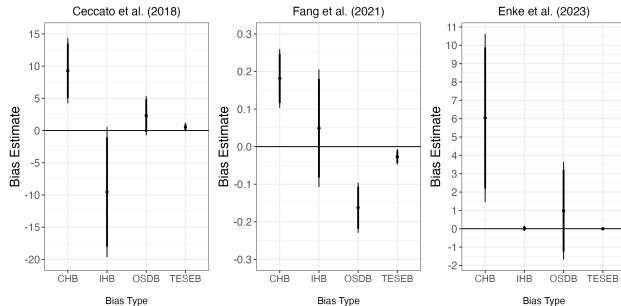
# OSDB and TESEB Can Have Opposite Signs



In this example,  $OSDB = 0.191$ , but  $TESEB = -0.038$

► **Takeaway:** Hypothetical biases on outcomes  $\neq$  hypothetical biases on TEs!

# Data



I re-analyze three recent hypothetical bias experiments which:

1. Vary *both* an intervention of interest and hypothetical stakes
2. Have publicly-available replication data

This allows me to directly compute and compare CHB w/ IHB and OSDB w/ TESEB

## Ceccato et al. (2018) (1/2)

Dictator game experiment where dictators must divide five-euro endowment between “Your Personal Envelope” and “Other Participant’s Envelope”

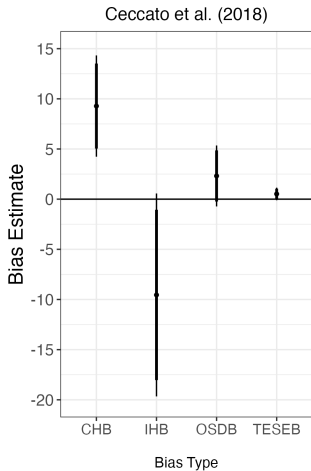
- ▶ Dictators randomized into real-stakes or hypothetical-stakes rooms, and thereafter randomized into seats

Dictators can receive one of two types of seats:

1. **‘Give’ framing**: Endowment initially stored in “Your Personal Envelope”
2. **‘Take’ framing**: Endowment initially stored in “Other Participant’s Envelope”

TE of interest is the effect of the ‘give’ framing treatment (relative to the ‘take’ framing control) on dictator transfers

## Ceccato et al. (2018) (2/2)



CHB and IHB exhibit opposite signs in this experiment

- ▶ Dictators give 9.3 p.p. *more* of their endowment when stakes are hypothetical
- ▶ However, effect of 'give' framing on endowment transfers is 9.5 p.p. *lower* when stakes are hypothetical

OSDB and TESEB share the same sign and statistical significance conclusions in this experiment

- ▶ However, OSDB is over four times the size of TESEB



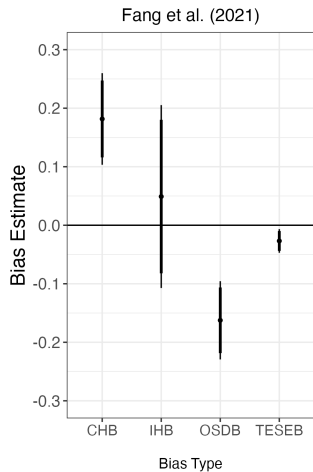
## Fang et al. (2021) (1/2)

Participants choose whether or not to buy a strawberry yogurt product under different price menus after being randomized into conditions where these decisions are made:

- ▶ With real/hypothetical stakes, and/or...
- ▶ In a virtual reality supermarket (vs. text-based/image-based control)

TE of interest is the effect of the virtual reality supermarket on purchase probability

## Fang et al. (2021) (2/2)



CHB and IHB yield completely different conclusions here

- ▶ Hypothetical stakes increase purchase probabilities by 18 p.p.
- ▶ The IHB on the virtual reality TE is  $< \frac{1}{3}$  the size of the CHB, and is not stat. sig. diff. from zero

OSDB and TESEB are both significantly negative in this experiment!

- ▶ However, the TESEB estimate is 83.5% smaller than the OSDB estimate
- ▶ The 13.6 p.p. difference between OSDB and TESEB is highly significant ( $SE = 2.9$  p.p.)

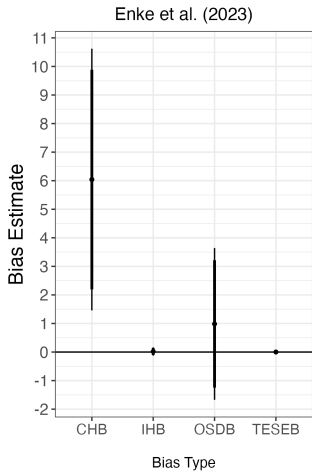
## Enke et al. (2023) (1/2)

Participants completing an anchoring task receive a personal anchor from 0-100 constructed from their birth year and phone number

- ▶ They then answer numerical trivia questions with correct answers from 0-100
- ▶ Participants are first primed with their anchor: 'Is the correct answer more or less than your anchor?'
- ▶ Participants then give an exact numerical answer
- ▶ This task is performed twice, first with no real stakes at play and thereafter with (probabilistically) real monetary stakes

TE of interest is the impact of the 0-100 anchor on numerical answers

## Enke et al. (2023) (2/2)



TE-relevant biases are way smaller than TE-irrelevant biases

- ▶ CHB is stat. sig.; IHB is not, and is 99.7% smaller than CHB
- ▶ Neither OSDB nor TESEB are stat. sig., but IHB is 99.8% smaller than CHB

This is likely due to scale effects

- ▶ Impacts of a one-point increase in a 0-100 anchor scale differently than binary switches between real and hypothetical stakes

**Intuition:** Hypothetical stakes won't affect all possible TEs on a given outcome in the exact same way!

## Practical Implications

Recent studies showing negligible CHB offer a clear practical implication: *Incentivization doesn't matter for this outcome, so you don't need to incentivize this outcome*

- ▶ This is not the right inference to make, especially when treatment effects are of interest

Ruling out meaningful hypothetical bias for all interventions targeting an outcome requires factorial designs that examine IHB/TESEB for every treatment on that outcome

- ▶ This is not a practically feasible research agenda, and stat. insig. IHB/TESEB estimates are not evidence of negligible IHB/TESEB

**Norms on (probabilistic) incentivization are there for a good reason** – it is untenable to assume that all treatment effects are unaffected by real stakes

- ▶ The recent rise of studies showing negligible CHB should thus not be interpreted as a justification to omit real stakes in intervention experiments

# Thank You For Your Attention!





Tinbergen Institute Discussion Paper




**Website:** <https://jack-fitzgerald.github.io/>

**Email:** [j.f.fitzgerald@vu.nl](mailto:j.f.fitzgerald@vu.nl)

# References I



-  Alesina, A., S. Stantcheva, and E. Teso (2018).  
Intergenerational mobility and preferences for redistribution.  
*American Economic Review* 108(2), 521–554.
-  Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström (2008).  
Eliciting risk and time preferences.  
*Econometrica* 76(3), 583–618.
-  Anderson, L. R., B. A. Freeborn, P. McAlvanah, and A. Turscak (2023).  
Pay every subject or pay only some?  
*Journal of Risk and Uncertainty* 66(2), 161–188.
-  Bardsley, N., R. Cubitt, G. Loomes, C. Starmer, R. Sugden, and P. Moffat (2009).  
*Incentives in Experiments* (1 ed.), pp. 244–285.  
Princeton University Press.

## References II

-  [Becker, G. M., M. H. Degroot, and J. Marschak \(1964\).](#)  
Measuring utility by a single-response sequential method.  
*Behavioral Science* 9(3), 226–232.
-  [Brañas-Garza, P., D. Jorrat, A. M. Espín, and A. Sánchez \(2022\).](#)  
Paid and hypothetical time preferences are the same: Lab, field and online evidence.  
*Experimental Economics* 26(2), 412–434.
-  [Camerer, C. F. and R. M. Hogarth \(1999\).](#)  
The effects of financial incentives in experiments: A review and capital-labor-production framework.  
*Journal of Risk and Uncertainty* 19(1/3), 7–42.



## References III

-  Clot, S., G. Grolleau, and L. Ibanez (2018).  
Shall we pay all? An experimental test of random incentivized systems.  
*Journal of Behavioral and Experimental Economics* 73, 93–98.
-  Cummings, R., S. Elliott, G. Harrison, and J. Murphy (1997).  
Are hypothetical referenda incentive compatible?  
*Journal of Political Economy* 105(3), 609–621.
-  Enke, B., U. Gneezy, B. Hall, D. Martin, V. Nelidov, T. Offerman, and J. van de Ven (2023).  
Cognitive biases: Mistakes or missing stakes?  
*Review of Economics and Statistics*, 1–15.
-  Fong, C. M. and E. F. Luttmer (2009).  
What determines giving to Hurricane Katrina victims? Experimental evidence on racial group loyalty.  
*American Economic Journal: Applied Economics* 1(2), 64–87.

## References IV



Guala, F. (2001).

Clear-cut designs versus the uniformity of experimental practice.

*Behavioral and Brain Sciences* 24(3), 412–413.



Hackethal, A., M. Kirchler, C. Laudenbach, M. Rizen, and A. Weber (2023).

On the role of monetary incentives in risk preference elicitation experiments.

*Journal of Risk and Uncertainty* 66(2), 189–213.



Harrison, G. W., E. Johnson, M. M. Mcinnes, and E. E. Rutström (2005).

Risk aversion and incentive effects: Comment.

*American Economic Review* 95(3), 897–901.






Hausman, J. (2012).





Contingent valuation: From dubious to hopeless.

*Journal of Economic Perspectives* 26(4), 43–56.





## References V

-  Kuziemko, I., M. I. Norton, E. Saez, and S. Stantcheva (2015).  
How elastic are preferences for redistribution? Evidence from randomized survey experiments.  
*American Economic Review* 105(4), 1478–1508.
-  Lane, T., D. Nosenzo, and S. Sonderegger (2023).  
Law and norms: Empirical evidence.  
*American Economic Review* 113(5), 1255–1293.
-  List, J. A. (2001).  
Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sportscards.  
*American Economic Review* 91(5), 1498–1507.



## References VI

-  Loomis, J. (2011).  
What's to know about hypothetical bias in stated preference valuation studies?  
*Journal of Economic Surveys* 25(2), 363–370.
-  Matousek, J., T. Havranek, and Z. Irsova (2022).  
Individual discount rates: A meta-analysis of experimental evidence.  
*Experimental Economics* 25(1), 318–358.
-  Murphy, J. J., P. G. Allen, T. H. Stevens, and D. Weatherhead (2005).  
A meta-analysis of hypothetical bias in stated preference valuation.  
*Environmental & Resource Economics* 30(3), 313–325.
-  Ortman, A. (2016).  
*Episodes from the early history of experimentation in economics*, pp. 195–217.  
Springer.

## References VII

-  Sefton, M. (1992).  
Incentives in simple bargaining games.  
*Journal of Economic Psychology* 13(2), 263–276.
-  Smith, V. L. (1976).  
Experimental economics: Induced value theory.  
*American Economic Review* 66(2), 274–279.
-  Smith, V. L. (1982, Dec).  
Microeconomic systems as an experimental science.  
*American Economic Review* 72(5), 923–955.
-  Sommet, N., D. L. Weissman, N. Cheutin, and A. Elliot (2022, Apr).  
How many participants do I need to test an interaction? Conducting an appropriate power analysis and achieving sufficient power to detect an interaction.  
*OSF Preprints*.

## References VIII

-  Svorencik, A. and H. Maas (2016).  
*The making of experimental economics: Witness seminar on the emergence of a field.*  
Springer.
-  Voslinsky, A. and O. H. Azar (2021).  
Incentives in experimental economics.  
*Journal of Behavioral and Experimental Economics* 93, 101706.