Identifying the Impact of Hypothetical Incentives on Experimental Outcomes and Treatment Effects

Jack Fitzgerald[*]

September 23, 2024

Abstract

Recent findings showing that outcome variables do not statistically significantly differ between incentivized and unincentivized conditions in experiments have spurred methodological challenges to experimental economics' disciplinary norm that experimental choices should be incentivized with real stakes. I show that the classical hypothetical bias measures estimated in these studies do not econometrically identify the kinds of hypothetical bias that matter in most modern experiments. Specifically, classical hypothetical bias measures are fully informative in 'elicitation experiments' where the researcher is uninterested in treatment effects (TEs). However, in 'intervention experiments' where TEs are of interest, classical measures are uninformative, and incentivization schemes matter if and only if TEs are heterogeneous between incentivization schemes. I demonstrate that classical hypothetical bias metrics can be misleading measures of hypothetical bias for intervention experiments, both econometrically and through several empirical applications. The fact that a given experimental outcome does not statistically significantly differ on average between incentivization conditions does not imply that all TEs on that outcome are unimpacted by incentivization. Therefore, the recent hypothetical bias literature does not justify leaving most modern experiments unincentivized. Norms in favor of completely or probabilistically incentivizing experimental choices remain useful for ensuring externally valid TEs in experimental economics.

Keywords: Interaction effects, meta-analysis, generalizability, bootstrap. JEL: C18, C90, D91.

[*] Affiliation: Vrije Universiteit Amsterdam; School of Business and Economics, Department of Ethics, Governance, and Society; Amsterdam, Noord-Holland, The Netherlands. Email: j.f.fitzgerald@vu.nl. ORCID: 0000-0002-0322-5104. Address: De Boelelaan 1105, Amsterdam, 1081 HV, The Netherlands. I am grateful to Maria Bigoni, Katharina Brütt, Florian H. Schneider, Jeroen van de Ven, and Linh Vu, as well as seminar, workshop, and conference participants from University of Amsterdam, University of Copenhagen, Vrije Universiteit Amsterdam, and the 2024 European Meeting of the Economic Science Association for valuable input on this paper. I gratefully acknowledge funding support from the Amsterdam Law and Behavior Institute. Replication code and directions for downloading necessary replication data can be found at https://osf.io/fe6jn/.

1 Introduction

Providing real incentives for experimental choices is one of the defining methodological hallmarks of experimental economics. It is a well-established norm in experimental economics that experimental choices should be incentivized, as theory holds that participants' motivation to maximize their real earnings can improve the generalizability of experimental behavior by overpowering other biases known to emerge in experimental environments (see Smith 1976; Smith 1982; Roth 1995; Camerer & Hogarth 1999; Hertwig & Ortmann 2001; Bardsley et al. 2009; Charness, Gneezy, & Halladay 2016; Svorenčík & Maas 2016; Clot, Grolleau, & Ibanez 2018). However, this norm is seeing a recent shift. Top economics publications are becoming increasingly receptive to publishing results from experiments with hypothetically-incentivized outcomes (e.g., see Golsteyn, Grönqvist, & Lindahl 2014; Cadena & Keys 2015; Kuziekmo et al. 2015; Alesina, Stantcheva, & Teso 2018; Sunde et al. 2022). A recent wave of research also shows that several commonly-elicited experimental outcomes do not statistically significantly differ on average between real incentive and hypothetical incentive conditions (see Matsouek et al. 2022; Brañas-Garza et al. 2023; Enke et al. 2023; Hackethal et al. 2023). Citing some of this recent hypothetical bias work (in particular Matsouek et al. 2022), the announcement for *Experimental Economics*' special issue on incentivization proclaims: "There is good rationale for incentivized experiments, but recently there has been evidence that incentivization may not always matter."[1]

This paper shows econometrically and empirically that the extant hypothetical bias literature does not statistically justify omitting real incentivization in most modern experiments. I begin by distinguishing 'elicitation experiments', where no intervention is varied and thus treatment effects (TEs) are not of interest, from 'intervention experiments', where at least one intervention is varied and its TE is of interest. Elicitation experiments dominated early experimental economics research, and though they remain important to this day, most modern experimental economics research is comprised of intervention experiments where researchers are interested in the TE(s) of some intervention(s).

Econometrically, classical tests for hypothetical bias – which are the primary tests em-

[1] See https://link.springer.com/journal/10683/updates/26740876. Accessed on 19 September 2024.

ployed in recent hypothetical bias studies – do not identify the hypothetical biases that matter for intervention experiments. Traditional hypothetical bias experiments randomize participants into either a real incentives or hypothetical incentives condition, elicit an outcome, and then test whether that outcome statistically significantly differs on average between incentivization conditions. This classical hypothetical bias measure is irrelevant for intervention experiments. The hypothetical bias relevant for intervention experiments is an interaction effect between the incentivization scheme and the treatment of interest. In contrast, classical methods for measuring hypothetical bias identify the average marginal effect of hypothetical incentivization on the outcome of interest, which has no general relationship with the relevant interaction effect between hypothetical incentivization and any treatment of interest. This is intuitive for two reasons. First, a researcher can not identify an interaction effect if all the researcher knows is the average marginal effect of one of the two variables in the interaction. Second, one should not expect that hypothetical incentivization will impact every possible intervention's TE on a given outcome variable in the exact same way.

Empirically, TE-irrelevant hypothetical bias measures often misidentify TE-relevant hypothetical biases in meaningful ways. I re-analyze the replication data of three recent hypothetical bias experiments that vary both a treatment of interest and hypothetical incentivization. These experiments permit me to directly estimate the interaction effects between hypothetical incentivization and treatments of interest, and to compare those interaction effects with the TE-irrelevant hypothetical bias estimates typically produced in the historical hypothetical bias literature. I show that TE-irrelevant hypothetical bias measures frequently yield different conclusions than TE-relevant hypothetical bias measures. In some cases, TE-irrelevant hypothetical bias measures can even exhibit sign flips when compared to TE-relevant hypothetical bias measures. That is, TE-irrelevant hypothetical bias estimates may be positive even when TE-relevant hypothetical biases are negative (and vice versa).

These findings cast doubt on the credibility and usefulness of recent advances in the hypothetical bias literature. Principally, my econometric results show that recent hypothetical bias findings – which show that certain outcomes do not statistically significantly differ between incentivization conditions – do not justify the general claim that real incentives 'do not matter' for all treatment effects on those outcomes. In fact, my empirical findings suggest that these recent results may plausibly mislead researchers who act on these findings by omitting real incentives in their intervention experiments, which may subsequently induce meaningful hypothetical biases that could threaten conclusions concerning TEs. Because ruling out hypothetical biases for a given intervention's TE on a given outcome functionally requires a factorial hypothetical bias experiment on that specific outcome and intervention, I also argue that it is unproductive and uninformative to conduct hypothetical bias experiments with the goal of 'paving the way' for future researchers to omit real incentives for their experimental choices. To that end, it remains useful to maintain existing norms in experimental economics that favor incentivizing experimental choices, and it is likely productive to augment these norms by explicitly permitting probabilistic incentivization in economic experiments as well.

This paper is structured as follows. Section 2 provides a taxonomy of experiments that clarifies the relevant differences between elicitation experiments and intervention experiments, and establishes notation for the paper. Section 3 discusses how hypothetical bias is measured in the historical literature, and Section 4 establishes econometrically why these classical methods for measuring hypothetical bias do not identify TE-relevant hypothetical biases. Section 5 provides three empirical applications demonstrating the differences between TE-relevant and TE-irrelevant hypothetical bias measures. Section 6 details the implications of my findings for the future of hypothetical bias research, and Section 7 concludes with a discussion on norms and best practices for future experimental economics research.

2 Terminology and Notation

I begin by establishing a simple taxonomy of experiments. For what follows, let $Y_i \in \mathbb{R}$ be the outcome variable of interest, and let $D_i \in \{0, 1\}$ be an experimental intervention of interest. For the purposes of this paper, my use of the word 'incentives' specifically refers to 'task-related incentives', which are payoffs that depend on participants' experimental choices (Bardsley et al. 2009).

I distinguish here between two types of experiments. The first type of experiment is an 'elicitation experiment.' This sort of experiment does not employ any intervention, and there are no TEs to estimate. The sole goal of an elicitation experiment is to use experimental procedures to elicit descriptive statistics concerning Y_i, usually sample means or medians. For example, a researcher interested in learning the average consumer's willingness to pay for a product may run an experiment employing a Becker-DeGroot-Marschak procedure to obtain an incentive-compatible measure of participants' willingness to pay (Becker, DeGroot, & Marschak 1964). This is undoubtedly an experiment, but there is no TE at play; the researcher is just interested in descriptive statistics on willingness to pay. This is thus an elicitation experiment.

The second type of experiment is an 'intervention experiment.' In contrast to an elicitation experiment, an intervention experiment employs an intervention of interest D_i, and the researcher is interested in the TE of this intervention. To extend the example from the previous paragraph, suppose that the researcher is interested in the effect of a particular product characteristic on willingness to pay. They could repeat the same Becker-DeGroot-Marschak experiment, but randomly assign half the participants to consider a product with the characteristic of interest, taking the difference in average willingness to pay between the two halves of the sample to estimate the TE of that product characteristic on willingness to pay. This would be an intervention experiment. The researcher may still be interested in descriptive statistics concerning Y_i in an intervention experiment. For instance, the experiment described in this paragraph is still an intervention experiment even if the researcher is additionally interested in estimating the mean willingness to pay within each treatment condition. So long as the experiment employs an intervention whose TE is of interest to the researcher, it is an intervention experiment.

In general, 'hypothetical bias' can be defined as the difference in the statistic of interest induced by a change in 'incentivization scheme' S_i, which is parameterized here as a dummy variable indicating that participant i receives real payoffs with probability p' rather than probability p. That is, for $p, p' \in [0,1]$ with $p \neq p'$,

$$S_i = \begin{cases} 0 \text{ if participant } i\text{'s payoffs are real with probability } p \\ 1 \text{ if participant } i\text{'s payoffs are real with probability } p' \end{cases} . \tag{1}$$

Ordinarily, $p = 1$ and $p' = 0$, so $S_i = 1$ indicates pure hypothetical incentives while $S_i = 0$ indicates pure real incentives. I use this definition of S_i throughout the remainder of this paper for simplicity, but this framework can be generalized to examine potential biases arising from switching between any pair of incentivization probabilities. This makes the statistical framework I introduce throughout this paper generalizable, permitting examination of not just the impacts of pure hypothetical incentivization, but also of probabilistic incentivization schemes. The specific bias induced when switching between incentivization schemes depends on the statistic of interest.

3 Historical Measurement of Hypothetical Bias

Many early seminal contributions in experimental economics are elicitation experiments. Early economic experiments are heavily focused on testing the predictions of prevailing economic theories, as well as documenting empirical regularities observed in laboratory experiments (Roth 1995). This was largely done using elicitation experiments that elicited diverse economic preferences and behaviors, including indifference curves for different bundles of goods (Thurstone 1931; Rousseas & Hart 1951), risk and ambiguity preferences (Allais 1953; Mosteller 1953), strategies in games (Flood 1958), and prices in experimental markets (Chamberlin 1948).

This history is important because the preponderance of elicitation experiments in the early years of experimental economics greatly influenced which statistical parameters experimental economists were interested in when disciplinary norms on experimental incentivization first emerged. Indeed, experimental economics largely already developed norms on providing real rather than hypothetical incentives for experimental tasks by the end of the 1950s (Roth 1995). The fact that early experimental economists were often more interested in descriptive statistics about people's basic economic preferences than the TEs of economically-relevant interventions influenced the reasons why experimental economists cared about real incentives, as well as the ways in which they measured bias when these incentives were not provided.

Two key rationales for experimental incentivization emerged in this early literature. First, experimental economists postulate that incentivization may impact the average preference or behavior elicited from a sample.[2] This implies that hypothetical incentives induce a bias on the expected value of Y_i. I term this bias 'classical hypothetical bias (CHB)', which can be written as

$$\text{CHB} \equiv \mathbb{E}\left[Y_i(p') - Y_i(p)\right]. \tag{2}$$

When the statistic of interest is the sample mean of Y_i, this bias can be easily parameterized in a linear model of the form

$$Y_i = \alpha + \delta S_i + \epsilon_i, \tag{3}$$

where $\text{CHB} = \delta$.

CHB is widely documented to be a substantial presence in economic experiments. Systematic evidence of CHB has existed since at least Camerer & Hogarth (1999), who document 36 studies in which zero-incentive conditions are assigned as an unconfounded treatment with a real-incentive control;[3] 26 of these studies (72%) show that hypothetical incentives have impacts on at least one outcome's central tendency. CHB is documented in a variety of experimental settings, including in ultimatum games (Sefton 1992), public goods games (Cummings et al. 1997), auctions (List 2001), and multiple price lists (Harrison et al. 2005). CHB is particularly severe in contingent valuation; experimental participants routinely overstate willingness to pay for public goods such as environmental services (see Hausman 2012). Meta-analytic estimates of CHB in contingent valuation range from 35% (Murphy et al. 2005) to 200% (List 2001). Even though a few recent experiments show small or nonexistent effects of incentivization on experimental outcomes (Matsouek et al. 2022; Brañas-Garza et

[2] These conjectures emerged early in experimental economics' history; see Roth (1995) for discussion of the 'Wallis-Friedman critique', which criticized elicitations of indifference curves using hypothetical choice menus and had a marked influence on leading experimental economists' decisions to incentivize their experiments (Svorenčík & Maas 2016).

[3] This is a subset of the full 74 experiments reviewed by Camerer & Hogarth (1999), specifically those that induce a '0 vs. L' treatment, or a '0' treatment with some more-incentivized control. My list excludes Scott, Farh, & Podsakoff (1988), as Camerer & Hogarth (1999) note that participants were not informed about incentives prior to the start of the experiment and that participants were in fact 'surprised' to receive their incentive payments at the end of the experiment.

al. 2023; Enke et al. 2023; Hackethal et al. 2023), there is a large body of work showing substantial risks of CHB in many experimental applications.

The second key rationale for experimental incentivization is noise reduction. Advocates of experimental incentivization contend that participants motivated by real incentives make choices more carefully and deliberatively than participants whose choices are unincentivized, and thus that incentivization reduces noise in experimental outcomes (Bardsley et al. 2009). Camerer & Hogarth (1999) note nine experiments where hypothetical incentives induce changes in the variance or convergence of experimental outcomes (usually increases in variance and decreases in convergence), while Hertwig & Ortmann (2001) find an additional two experiments that exhibit such effects.

However, the manner in which these 'noise reduction' effects are measured greatly varies across studies. Some studies simply discuss changes in the standard deviation (SD) or variance of an outcome between incentivization conditions (see for example Wright & Anderson 1989; Ashton 1990; Irwin, McClelland, & Schulze 1992; Forsythe et al. 1994). Others judge noise by outcome deviations from some predicted value, such as price deviations from a competitive market price (see Edwards 1953; Smith 1962; Smith 1965; Jamal & Sunder 1991; Smith & Walker 1993).

There is also a concern about the precision of these variance differences. Incentivization-driven changes in variance (either around the sample mean of Y_i or around some theory-predicted outcome value) are virtually always reported in a purely descriptive fashion, with no measure of precision, such as a standard error (SE), to qualify the magnitude of between-condition changes in variance.[4] It is thus unclear whether existing estimates of incentivization's impact on noise reflect true effects or arise as a simple artefact of sampling variation. This is especially true given that noise effects are typically a second-order concern of hypothetical bias research and are thus not usually systematically investigated.

For the purposes of this paper, I parameterize the effect of hypothetical incentivization

[4] Recent attempts to qualify the significance of differences in variance between incentivization conditions often approach this task using non-parametric approaches such as Kolmogorov-Smirnov tests (e.g., see Brañas-Garza et al. 2023; Hackethal et al. 2023). However, such non-parametric tests only identify significant differences in *distributions*, which are defined not just by parameter variances, but also by centrality measures and other moments.

on noise as an ‘outcome SD bias (OSDB)’, which can be written as

$$\text{OSDB} \equiv \mathbb{E}\left[\sigma_{Y_i}(p') - \sigma_{Y_i}(p)\right]. \tag{4}$$

A point estimate of this bias can be obtained by simply taking the difference in outcome SD σ_{Y_i} between incentivization conditions; an SE of this estimate can be obtained via bootstrap (see Section 5 for further details). I define noise in this way because not all experimental outcomes have clear predicted values that ‘should’ be observed in experimental data.

Many recent studies concerning the effects of incentivization on experimental results are exclusively analyzing CHB and OSDB. Matsouek et al. (2022) meta-analytically find that the average individual discount rate under incentivized conditions does not statistically significantly differ from the average discount rate found under unincentivized conditions. Brañas-Garza et al. (2023) show that the means and SDs of time discounting factors are not statistically significantly different between paid and hypothetical conditions, while Hackethal et al. (2023) find that the same is true of the number of risky choices that participants make in a multiple price list experiment. Enke et al. (2023) find that the mean differences in correct answers on the cognitive reflection test, a base rate neglect test, and a contingent reasoning test are not statistically significantly different between unincentivized and incentivized conditions. These studies are reporting estimates of CHB, and both Brañas-Garza et al. (2023) and Hackethal et al. (2023) are additionally reporting evidence on OSDB.

Though CHB and OSDB are fully informative measures of hypothetical bias in elicitation experiments, which dominated the landscape of experimental economics when norms on incentivization first emerged, most modern work in experimental economics (and experimental social sciences in general) is not exclusively limited to elicitation experiments. Charness & Pingle (2022) identify 20 top papers in the history of experimental economics based on citations, expert recommendations, and voting, 14 of which are published after 1990 and 10 of which are published after 2000.[5] The only clear elicitation experiment in this collection is Andreoni & Miller (2002); the remainder vary at least one intervention whose TE is of interest to the article.

[5] A notable omission from this collection is Kahneman & Tversky (1979), which is also an intervention experiment.

Although elicitation experiments remain important to this day, many researchers have less interest in descriptive statistics and more interest in the clean causal TEs that can be obtained from intervention experiments. Such experimental TEs were, and still are, crucial antecedents of the credibility revolution in economics (Angrist & Pischke 2010). However, as the next section shows, CHB and OSDB are completely uninformative measures of hypothetical bias when TEs are of interest.

4 Hypothetical Bias for Treatment Effects

4.1 Treatment Effect Point Estimates: IHB

CHB is irrelevant for describing hypothetical bias on TEs. In fact, Equation 3 shows that CHB can be modeled and estimated without any regard for intervention D_i. Any framework used to identify the effect of incentivization on TEs must incorporate D_i, and must allow the possibility that TEs are impacted by incentivization scheme S_i.

I consider a simple 2x2 factorial design where treatment D_i and incentivization scheme S_i are both randomized with equal weight across participants. As in Guala (2001), I model the effects of D_i and S_i in a simple heterogeneous treatment effects framework:

$$Y_i = \alpha + \beta_1 D_i + \beta_2 S_i + \beta_3 (D_i \times S_i) + \mu_i. \tag{5}$$

Randomization of D_i and S_i confers unconfoundedness: $\mathbb{E}\left[\mu_i | D_i, S_i\right] = 0$. Potential outcomes can be modelled as

$$\tau_i = Y_i(1, S) - Y_i(0, S) = \begin{cases} \beta_1 \text{ if } S_i = 0 \\ \\ \beta_1 + \beta_3 \text{ if } S_i = 1 \end{cases}, \tag{6}$$

with $Y_i(D, S)$ representing the potential outcome of Y_i depending on intervention status D and incentivization scheme status S. For what follows, suppose that the statistic of interest to the researcher is TE $\tau \equiv \mathbb{E}\left[\tau_i\right]$.

The hypothetical bias on the point estimate of τ can be derived as a simple difference in differences, which I term 'interactive hypothetical bias (IHB)':

$$\begin{aligned} \text{IHB} &\equiv \mathbb{E}\left[\tau_i\left(p'\right) - \tau_i\left(p\right)\right] && (7) \\ &= \mathbb{E}\left[Y_i(1,1) - Y_i(0,1)\right] - \mathbb{E}\left[Y_i(1,0) - Y_i(0,0)\right] \\ &= (\beta_1 + \beta_3) - \beta_1 = \beta_3. && (8) \end{aligned}$$

Hypothetical incentivization thus biases a TE's point estimate if and only if $\beta_3 \neq 0$. This yields an intuitive conclusion: under the data-generating process in Equation 5, hypothetical bias on the point estimate of an intervention's TE can be fully identified as the interaction effect between that intervention and the incentivization scheme.

IHB is a fully informative measure of hypothetical bias in intervention experiments, but CHB does not identify this term. Under the data-generating process in Equation 5, CHB is the marginal effect of S_i on Y_i:

$$\delta_i = Y_i\left(D,1\right) - Y_i\left(D,0\right) = \begin{cases} \beta_2 \text{ if } D_i = 0 \\ \beta_2 + \beta_3 \text{ if } D_i = 1 \end{cases}. \qquad (9)$$

Given the aforementioned unconfoundedness yielded by randomization, the CHB estimated from a simple regression specification of Equation 3 will return $\hat{\delta} = \mathbb{E}\left[\delta_i\right]$, which by Equation 9 can be decomposed as

$$\mathbb{E}\left[\delta_i\right] = \beta_2 + \mathbb{E}\left[D_i\right]\beta_3. \qquad (10)$$

CHB thus does not identify IHB, and inferring IHB from CHB can yield misleading conclusions. Per Equation 10, if $|\beta_2|$ is large and $\beta_3 = 0$, then CHB will be large while IHB will be zero. In a similar manner, if $\beta_2 = -\mathbb{E}\left[D_i\right]\beta_3$, then CHB will be zero for arbitrarily large IHB. In fact, CHB and IHB differ almost always, as under Equations 8 and 10, CHB and IHB differ whenever it holds that $\beta_3 \neq \mathbb{E}\left[\delta_i\right]$, which holds whenever $\beta_3 \neq \frac{\beta_2}{1-\mathbb{E}[D_i]}$. This last condition holds almost always, which is intuitive, as the interaction effect between an intervention and some moderator virtually never perfectly positively correlates with the average marginal effect of the moderator itself.

Recent research on hypothetical bias in experiments – which focuses almost exclusively on CHB – must be understood in this context. Though Matsouek et al. (2022), Brañas-Garza et al. (2023), and Hackethal et al. (2023) respectively find no statistically significant CHBs on discount rates, time preferences, and risk preferences, this in no way implies that incentivization has zero impact for all intervention TEs on these outcomes. To identify IHBs for a given intervention D_i and a given outcome Y_i, one must vary both D_i and S_i in an experimental setting that permits unconfounded estimation of both the individual and joint impacts of D_i and S_i on Y_i. Identification of IHBs is thus not possible in experiments that only vary S_i. Further, there is no 'one true' IHB for all interventions of interest, as different interventions likely exhibit different IHBs for the same outcome.

4.2 Treatment Effect Standard Errors: TESEB

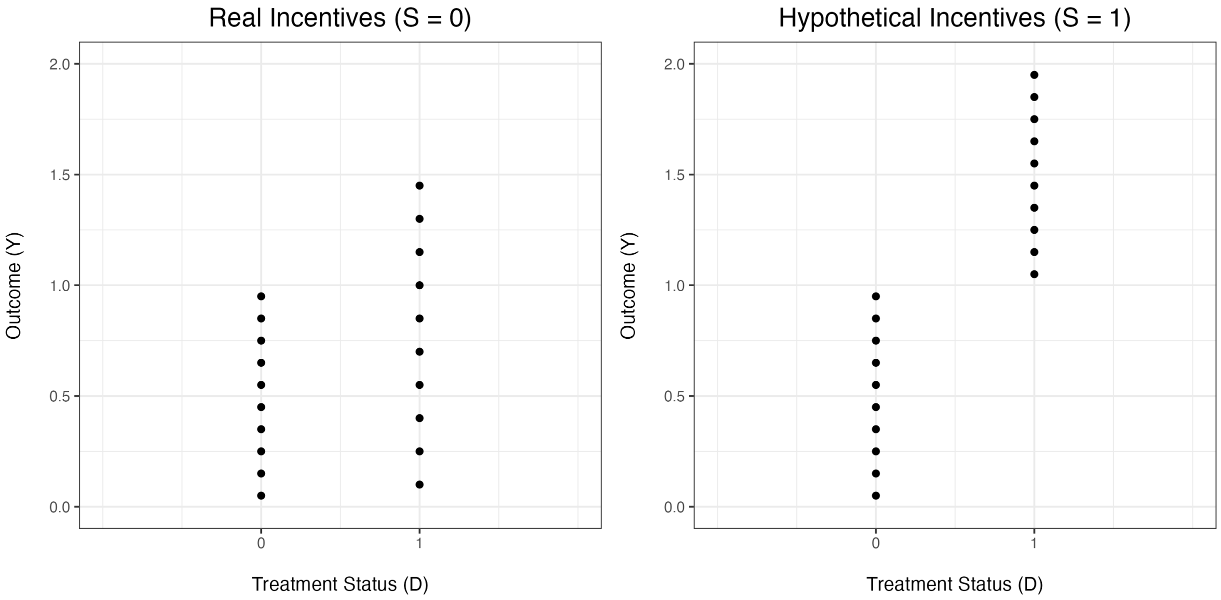
Hypothetical bias on TE SEs can be identified in a similar fashion to Equation 7. I parameterize hypothetical bias on TE precision as a 'TE SE bias (TESEB)':

$$\text{TESEB} \equiv \text{SE}\left(\tau\left(p'\right)\right) - \text{SE}\left(\tau\left(p\right)\right). \tag{11}$$

In practice, point estimates for TESEBs can be obtained by taking the differences in TE SEs between incentivization conditions, and SEs for TESEBs can be estimated via bootstrapping.

OSDB does not identify this bias. The best way to show this is through a simple counterexample where OSDB and TESEB exhibit opposite signs. Figure 1 displays data points from two simulated datasets of 20 observations each, where for both datasets, $D_i = 0$ for $i \in \{1, 2, \cdots 10\}$ and $D_i = 1$ for $i \in \{11, 12, \cdots 20\}$. The first dataset is constructed using the data-generating process

$$Y_i = \begin{cases} 0.05 + 0.1(i-1) \text{ if } i \in \{1, 2, \cdots 10\} \;\; (D_i = 0) \\ -0.05 + 0.15(i-10) \text{ if } i \in \{11, 12, \cdots 20\} \;\; (D_i = 1) \end{cases}, \tag{12}$$



Note: The graphs plot data points from two simulated datasets. The left graph's data points arise from the data-generating process in Equation 12, while the right graph's data points arise from the data-generating process in Equation 13.

Figure 1: An Example Where OSDB and TESEB Hold Opposite Signs

and the second dataset is constructed using the data-generating process

$$Y_i = \begin{cases} 0.05 + 0.1(i-1) \text{ if } i \in \{1, 2, \cdots 10\} \;\; (D_i = 0) \\ 1.05 + 0.1(i-11) \text{ if } i \in \{11, 12, \cdots 20\} \;\; (D_i = 1) \end{cases} . \tag{13}$$

For purposes of exposition, suppose that these two datasets are two halves of an experimental dataset that randomizes assignment of D_i and S_i, where the first half (generated by the process in Equation 12) belongs to a sample facing real incentives (i.e., $S_i = 0$), whereas the second half (generated by the process in Equation 13) belongs to a sample facing hypothetical incentives (i.e., $S_i = 1$). It is clearly visible that the outcome SD in the sample facing hypothetical incentives (0.592) is higher than that in the sample facing real incentives (0.401), so OSDB is positive. However, the TE SE from a simple linear regression model of Y_i on D_i is smaller in the sample facing hypothetical incentives (0.135) than in the sample facing real incentives (0.173), so TESEB is negative.[6] This simulated example

[6] When HC3 heteroskedasticity-robust SEs are employed (see MacKinnon & White 1985), the TE SE in the sample facing hypothetical incentives (0.143) is still smaller than that in the sample facing real incentives (0.182).

demonstrates that findings on 'noise' based on OSDBs computed in experiments that only vary incentivization schemes can yield wildly misleading conclusions concerning the effect of incentivization on the precision of TEs.

4.3 Meta-Analytic Approaches

One method by which hypothetical bias researchers attempt to directly estimate IHB is by meta-analytically comparing the TEs of studies with and without real incentives. For instance, Li, Maniadis, & Sedikides (2021) conduct a meta-analysis of studies concerning anchoring effects on willingness to pay/accept. They find that the TEs of studies that use real incentives are not statistically significantly different from those that use hypothetical choice settings, and therefore conclude that financial incentives have no discernible effect on TEs. One could conceivably use a similar approach to estimate TESEBs by comparing meta-analytic averages of TE SEs under different incentivization conditions, though Li, Maniadis, & Sedikides (2021) do not make this comparison.

Meta-analyses like this do not offer clean causal estimates of the impacts of real incentivization, as the choice to incentivize an experiment is not random. The derivation of IHB as a simple interaction effect between treatment D_i and incentivization scheme S_i in Equation 8 depends on a joint unconfoundedness assumption over both the treatment the incentivization scheme, $\mathbb{E}\left[\mu_i | D_i, S_i\right] = 0$. This is readily achieved *within* a factorial experiment if both treatment status and incentivization scheme are randomly assigned. However, this unconfoundedness condition is not generally satisfied when comparing TEs *across* experiments, as experimental incentivization schemes are typically not randomly assigned, and are likely correlated with other factors that simultaneously impact TEs and their SEs.

One important dimension on which incentivization schemes are confounded is academic discipline. Naturally, some disciplines are more likely to offer real incentives than others, and these disciplines vary in important ways on dimensions including participant pools and procedural norms in experimentation (see Hertwig & Ortmann 2001). To fix a simple example, suppose that in a given meta-analytic dataset, all experiments employing real incentives are run by economists whereas all experiments providing hypothetical incentives are run by psychologists. Further, suppose that all of the economics experiments are run using pools

Article	Outcome	Treatment	CHB	IHB	OSDB	TESEB	N
Ceccato et al. (2018)	% of endowment transferred (0-100)	Give (vs. take) framing (0/1)	9.28 (2.575)	-9.547 (5.16)	2.312 (1.55)	-2.39 (1.423)	348
Fang et al. (2021)	Purchasing yogurt (0/1)	Virtual reality (0/1)	0.182 (0.04)	0.049 (0.08)	-0.162 (0.034)	0.579 (1.439)	1024
Enke et al. (2023)	Answer (0-100)	Numerical anchor (0-100)	6.04 (2.338)	0.019 (0.074)	0.985 (1.356)	0.154 (2.052)	626

Note: CHB denotes 'classical hypothetical bias', IHB represents 'interactive hypothetical bias', OSDB denotes 'outcome standard deviation bias', TESEB denotes 'TE SE bias', and N is the effective sample size. SEs are presented in parentheses.

Table 1: Empirical Estimates of Hypothetical Bias Measures

of economics students, whereas all of the psychology experiments are run using pools of psychology students. In order to interpret the difference in TEs between these groups of experiments as a causal effect of hypothetical incentivization, one must be willing to assume (among other things) that economics students respond to treatment in the exact same way as psychology students. Even this simple assumption is untenable; economics students differ from psychology students in important ways, and the same treatment can impact economics students and psychology students in significantly different ways (Van Lange, Schippers, & Balliet 2011; van Andel, Tybur, & Van Lange 2016). Meta-analytic differences between TEs thus do not generally cleanly identify IHBs, and for similar reasons, meta-analytic differences between TE SEs do not generally identify TESEBs.

5 Empirical Applications

To show that historical hypothetical bias measures misidentify hypothetical biases on treatment effects empirically (and not just on paper), I leverage replication data on three recent hypothetical bias experiments to directly estimate CHB, IHB, OSDB, and TESEB in these experiments. Apart from data availability constraints, I select these experiments because they leverage factorial designs that manipulate both the incentivization scheme and another treatment whose TE is of interest. The results of my empirical analyses are presented in Table 1. Throughout the rest of this section, I overview each of the three experiments that I re-examine, detail how CHB, IHB, OSDB, and TESEB are computed for each experiment, and discuss the ways in which my results show that TE-irrelevant hypothetical bias measures misidentify TE-relevant hypothetical bias measures.

5.1 Ceccato et al. (2018)

Ceccato et al. (2018) report the results of an experiment conducted using participants who play double-anonymous dictator games. Participants are randomly assigned either into a room where incentives are hypothetical or into a room where incentives are real, after which participants are randomized into a given seat in their assigned room. Dictators are faced with two envelopes, one titled "Your Personal Envelope" and the other titled "Other Participant's Envelope", and must decide the allocation of a five-euro endowment between these two envelopes. Dictators can receive a seat with 'give' framing, where the endowment is initially stored in "Your Personal Envelope", or a seat with 'take' framing, where the endowment is initially stored in "Other Participant's Envelope". The experiment also takes steps to manipulate the gender of the dictator and the passive player, but for the purposes of this replication, I focus exclusively on the effect of the 'give' framing treatment (compared to the 'take' framing control) on dictator transfers. Replication data for the experiment reported in Ceccato et al. (2018) is provided by Schwieren et al. (2018).

For this experiment, I first compute IHB in an ordinary least squares model of the form

$$\%\text{Trans}_i = \alpha + \beta_1 \text{Give}_i + \beta_2 S_i + \beta_3 \text{Give}_i S_i + \mu_i,$$

where $\%\text{Trans}_i$ is the percentage of the endowment transferred by dictator i, Give_i indicates the 'give' framing treatment, S_i indicates hypothetical incentives, and β_3 is the IHB estimate of interest. From this model, I use the `avg_slopes()` command in the `marginaleffects` R suite to obtain CHB as the average marginal effect of S_i on $\%\text{Trans}_i$ (see Arel-Bundock, Greifer, & Bacher 2024). SEs for both CHB and IHB are computed using the HC3 variance-covariance estimator (see Hayes & Cai 2007).

I obtain a point estimate of OSDB by simply subtracting the within-sample SD of $\%\text{Trans}_i$ for observations with $S_i = 0$ from the same SD for observations with $S_i = 1$. I then run ordinary least squares models of the form

$$\%\text{Trans}_i = \alpha_H + \tau_H \text{Give}_i + \mu_i, S_i = 1$$

$$\%\text{Trans}_i = \alpha_R + \tau_R \text{Give}_i + \mu_i, S_i = 0.$$

That is, I separately regress %Trans$_i$ on Give$_i$ in the subsamples where $S_i = 1$ and $S_i = 0$ (respectively). My TESEB point estimate is simply $\text{SE}(\hat{\tau}_H) - \text{SE}(\hat{\tau}_R)$. To obtain SEs for the OSDB and the TESEB, I repeat my procedures for obtaining the OSDB and TESEB point estimates on 10,000 bootstrap samples; my SE estimates for OSDB and TESEB are respectively the SDs of the OSDBs and TESEBs from my bootstrap sample.

Table 1 shows that CHB and OSDB wildly misidentify IHB and TESEB (respectively) in Ceccato et al. (2018), with both TE-irrelevant hypothetical bias measures exhibiting sign flips when compared to their respective TE-relevant hypothetical bias measures. CHB is quite significantly positive, with hypothetical incentives causing dictators to transfer over nine percentage points more of their endowment to recipients. This is intuitive, as people tend to overstate their willingness to give when stakes are not real (e.g., see Sefton 1992). However, the IHB on the impact of 'give' framing on endowment transfers is *negative*, and is even larger in magnitude than the CHB on endowment transfers (though this IHB is quite imprecise). Turning to the precision-related hypothetical bias measures, OSDB and TESEB are both imprecise in this setting, but they take on opposite signs. The dispersion of endowment transfers is larger when incentives are hypothetical, but the SE of the TE of 'give' framing on endowment transfers is smaller when incentives are hypothetical.

5.2 Fang et al. (2021)

Fang et al. (2021) examine whether the use of virtual reality marketplaces can mitigate hypothetical bias in choice experiments. Participants are faced with the choice to purchase an original strawberry yogurt, a light strawberry yogurt, or neither of the two. Participants are randomized into one of five between-participant conditions. The first condition is a hypothetically-incentivized condition where participants make product choices based on photos of the products. In the second and third conditions, participants choose between products based on textual information, namely their nutritional labels; one of these two conditions is a hypothetical incentives condition while the other is a real incentives condition. In the fourth and fifth conditions, participants make product decisions in a virtual reality supermarket setting; as with the text-based treatments, one of these two conditions is truly incentivized while the other is hypothetically incentivized. Once randomized to a condition, each participant makes purchase decisions four times, each time facing a different price menu.

Because it is the primary target of the Fang et al. (2021) experiment, I focus specifically on the effect of virtual reality on the decision to purchase. I specifically estimate IHB in a panel data random effects model of the form

$$\text{Buy}_{i,p} = \alpha + \beta_1 \text{VR}_i + \beta_2 S_i + \beta_3 \text{VR}_i S_i + \mu_{i,p},$$

where i indexes the participant and p indexes the price menu. I code $\text{Buy}_{i,p}$ as a dummy indicating that participant i chooses to purchase either the original or light yogurt when facing price menu p, VR_i as a dummy indicating that participant i is facing one of the two virtual reality treatments, and S_i as a dummy indicating that participant i is facing one of the three conditions with hypothetical stakes. As for Ceccato et al. (2018), β_3 is the IHB parameter of interest, and I compute CHB using the `avg_slopes()` command in the `marginaleffects` R suite. SEs for both IHB and CHB are computed using an HC3 variance-covariance estimator, with SEs clustered at the participant level.

As in my re-analysis of Ceccato et al. (2018), I obtain a point estimate of OSDB by computing the difference in SDs of $\text{Buy}_{i,p}$ between the samples where $S_i = 1$ and $S_i = 0$. I run random effects panel data models of the form

$$\text{Buy}_{i,p} = \alpha_H + \tau_H \text{VR}_i + \mu_{i,p}, \;\; S_i = 1$$
$$\text{Buy}_{i,p} = \alpha_R + \tau_R \text{VR}_i + \mu_{i,p}, \;\; S_i = 0$$

and compute the TESEB point estimate as $\text{SE}(\hat{\tau}_H) - \text{SE}(\hat{\tau}_R)$. To estimate SEs for OSDB and TESEB, I repeat the procedures to obtain point estimates for OSDB and TESEB in 10,000 cluster bootstrap samples (where participants i, rather than rows $\{i, p\}$, are resampled with replacement). I respectively compute the SEs of OSDB and TESEB as the SDs of the OSDB and TESEB point estimates in my bootstrap sample.

Table 1 shows that TE-irrelevant hypothetical bias measures yield completely different conclusions than TE-relevant hypothetical bias measures in Fang et al. (2021). CHB is significantly positive in this experiment, with participants facing hypothetical stakes being over 18 percentage points more likely to choose to purchase one of the two yogurts than participants facing real stakes. This reflects the intuitive and well-documented fact that people routinely overstate their willingness to pay for products and services when stakes are hypothetical (see List 2001; Murphy et al. 2005; Hausman 2012). However, the IHB estimate in this experiment is less than one third the size of the CHB estimate, and is not statistically significantly different from zero. Turning to OSDB and TESEB, another sign flip arises. Hypothetical incentives appear to significantly decrease the dispersion of purchase decisions, decreasing the SD of $\text{Buy}_{i,p}$ by over 16 percentage points. However, the TESEB estimate is positive, and is roughly 3.5 times the size of the OSDB estimate. Despite its larger size, the TESEB estimate is too imprecise to yield confident statistical significance conclusions.

5.3 Enke et al. (2023)

As aforementioned, Enke et al. (2023) examine hypothetical biases for a variety of commonly-elicited experimental outcomes. Participants first complete two out of four possible tasks without any task-related incentives, and are thereafter randomized in between-participants fashion into either a low-stakes or high-stakes real incentives condition to repeat these same two tasks. For three of the four tasks, there are no interventions at play, and thus it is only possible to examine CHB and OSDB.[7] However, Enke et al. (2023) also examine the impact of incentivization in an anchoring context, where there is a clear TE at play (that is, the anchoring effect); it is possible to examine IHB and TESEB in this specific setting.

Participants facing the anchoring task are asked to answer two of four randomly-assigned numerical questions whose answers range from 0-100;[8] one of these questions is faced under incentivized conditions and the other is faced under unincentivized conditions. For a given anchoring question, participants are first primed with an anchor constructed using the first two digits of their birth year and the last digit of their phone number, faced with the question of whether the numerical answer to the question is greater than or less than the

[7] These outcomes include scores on the cognitive reflection test, answers for a base rate neglect question, and answers for a contingent reasoning question.

[8] Questions include "Is the time (in minutes) it takes for light to travel from the Sun to the planet Jupiter more than or less than ANCHOR minutes?" and "Is the population of Uzbekistan as of 2018 greater than or less than ANCHOR million?" See Appendix B.3 in Enke et al. (2023).

anchor. Participants must thereafter provide an exact numerical answer to the question. In the incentivized conditions, participants earn a bonus if their answer to the question is within two points of the correct answer. My replication of Enke et al. (2023) focuses only on the sample facing the anchoring task, and to get as close as possible to examining extensive margin effects of real vs. hypothetical incentivization, I exclude the portion of the sample subjected to the high-stakes incentivization treatment. Replication data for Enke et al. (2023) is provided by Enke et al. (2021).

Estimation procedures for Enke et al. (2023) closely mirror those for Fang et al. (2021). IHB is computed in a panel data random effects model of the form

$$\text{Answer}_{i,c} = \alpha + \beta_1 \text{Anchor}_i + \beta_2 S_c + \beta_3 \text{Anchor}_i S_c + \mu_{i,c},$$

where i indexes the participant and c indexes the condition. Here Anchor_i represents participant i's anchor and S_c is a dummy indicating that the participant is facing the no-stakes condition. I subsequently use the `avg_slopes()` command in the `marginaleffects` R suite to compute CHB as the average marginal effect of S_c on $\text{Answer}_{i,c}$. SEs for both IHB and CHB are computed using the HC3 variance-covariance estimator, with SEs clustered at the participant level.

The OSDB point estimate is computed as the SD of $\text{Answer}_{i,c}$ when $S_c = 1$ minus the SD of $\text{Answer}_{i,c}$ when $S_c = 0$. I then run random effects panel data models of the form

$$\text{Answer}_{i,c} = \alpha_H + \tau_H \text{Anchor}_i + \mu_{i,c}, \;\; S_c = 1$$
$$\text{Answer}_{i,c} = \alpha_R + \tau_R \text{Anchor}_i + \mu_{i,c}, \;\; S_c = 0$$

and obtain TESEB point estimate $\text{SE}(\hat{\tau}_H) - \text{SE}(\hat{\tau}_R)$. As for Fang et al. (2021), I then re-obtain the OSDB and TESEB point estimates in 10,000 cluster bootstrap samples. SEs of OSDB and TESEB are respectively computed as the SDs of the OSDB and TESEB point estimates in the bootstrap sample.

My replication of Enke et al. (2023) shows how TE-irrelevant hypothetical bias measures misidentify TE-relevant hypothetical bias not just in terms of qualitative conclusions, but also in terms of scale. The CHB estimate is statistically significantly different from zero; participants appear to offer numerical answers roughly six points higher (out of 100) when stakes are real. However, the IHB estimate is minuscule by comparison, and is not statistically significantly different from zero. This partially reflects the fact that real incentive provision and numerical anchors impact numerical answers at completely different scales. It is intuitive that a one-point increase in a 0-100 numerical anchor will have a relatively small impact on numerical answers compared to a binary switch from unincentivized to incentivized conditions. It is worth considering that similar scale differences emerge between CHB and IHB in many other applications. Likewise, the TESEB estimate is less than one sixth the size of the OSDB estimate, which may reflect similar differences in scale. However, both the OSDB and TESEB estimates are too imprecisely estimated to yield confident statistical significance conclusions.

6 Discussion

6.1 Practical Implications of Hypothetical Bias Research

The practical reason why a researcher would like to be able to use statistically insignificant CHBs to ‘rule out’ hypothetical bias for a given experimental outcome is clear: researchers would like to be able to run cheaper intervention experiments by omitting real incentives for experimental choices. To that end, some researchers point to hypothetical bias studies that report results on CHB for a given outcome variable as justification to not incentivize intervention experiments concerning that outcome. For example, Matsouek et al. (2022) and Brañas-Garza et al. (2023) find CHBs on time preferences that are not statistically significantly different from zero. One could point to the CHBs found in these studies as justification to not incentivize an intervention experiment where time preferences are the outcome of interest, reasoning that the statistically insignificant CHB estimates in Matsouek et al. (2022) and Brañas-Garza et al. (2023) are evidence that real incentives ‘do not matter’ for eliciting time preferences.

However, this interpretation is not justified. My identification results in Section 4 and my empirical results in Section 5 make clear that TE-irrelevant hypothetical bias measures (namely CHB and OSDB) can wildly misidentify TE-relevant hypothetical bias measures (specifically IHB and TESEB, respectively). Showing that CHB for a particular outcome is not statistically significantly different from zero does not imply that all (or any) treatments targeting that outcome will exhibit negligible IHB. For a researcher to be confident that omitting real incentives will have negligible effects on their experiment's TEs, they must have *a priori* knowledge that both IHB and TESEB will be negligible for every combination of all interventions and outcomes in their experiment. Given the lack of research on IHB and TESEB in the current literature, it is unlikely that researchers possess this knowledge *a priori* when running an unincentivized experiment.

6.2 Statistical (In)significance

Even if a researcher genuinely has evidence that all hypothetical biases of relevance to their experiment are not statistically significantly different from zero, this is still not credible evidence that hypothetical incentivization has negligible consequences. Much of the present hypothetical bias literature interprets *statistically insignificant* hypothetical bias estimates as evidence of *practically negligible* hypothetical bias. This is a widely-known misinterpretation of statistical (in)significance, which can yield high Type II error rates if applied generally (see Altman & Bland 1995; Wasserstein & Lazar 2016; Fitzgerald 2024). Further, statistically insignificant hypothetical biases can still meaningfully change experimental conclusions, as the difference between a statistically significant estimate and a statistically insignificant estimate is not itself statistically significant (Gelman & Stern 2006).

The Type II error rates incurred by misinterpreting statistically insignificant hypothetical bias as *ipso facto* evidence of practically negligible hypothetical bias are amplified for TE-relevant hypothetical bias measures, which tend to be considerably underpowered. For example, IHBs are interaction effects, which are notoriously imprecise and difficult to sufficiently power. In a simple heterogeneous treatment effects framework, if a main effect is sufficiently powered with N observations, and the interaction effect is half the size of the main effect, then it will take $8N$ observations to sufficiently power that interaction effect (Muralidharan, Romero, & Wüthrich 2023).

This property can be observed in my empirical results. For instance, Table 1 shows that the CHB on endowment transfers in Ceccato et al. (2018) is statistically significant, with a t-statistic exceeding 3.5. However, despite the fact that the IHB estimate for the framing effect on endowment transfers is larger than the CHB estimate, the IHB estimate is not statistically significant because its standard error is double that of the CHB estimate. If one is prepared to consider this experiment's CHB estimate to be practically significant, then one should not be simultaneously prepared to deem its IHB estimate as negligible simply because it is less precisely estimated. Additionally, SEs of OSDBs and TESEBs are quite imprecise in my replication results, providing some suggestive empirical evidence that similar power issues may emerge for OSDB and TESEB.

6.3 Non-Inferiority and Equivalence Testing Approaches

What would be credible evidence that an experimental TE is practically unaffected by hypothetical incentivization? What most researchers ultimately care about is whether the conclusions that they make about TEs are meaningfully impacted by incentivization schemes. In practice, this means that before a researcher chooses to omit real incentives, they should be certain that IHBs will be small enough that statistical significance conclusions concerning their TEs of interest will not change if their experiment omits real incentives.

When TEs of interest are not statistically significantly different from zero under real incentives conditions, credible evidence that IHBs are practically negligible can be obtained using equivalence testing. For instance, presume that a researcher conducts the factorial experiment that I devise in Section 4, such that both treatment D_i and incentivization scheme S_i are exogenously varied amongst participants i. Further, suppose that the TE estimate of interest under real incentive conditions, $\hat{\tau}(p)$, is not statistically significantly different from zero. However, suppose that if the point estimate for $\hat{\tau}(p)$ were to increase by $\epsilon_+ > 0$ or by $\epsilon_- < 0$, then $\hat{\tau}(p)$ would become statistically significantly different from zero.

Then one can use equivalence testing to test the following hypotheses:

$$
\begin{aligned}
&H_0 : \epsilon_- > \tau(p') - \tau(p) \text{ or } \tau(p') - \tau(p) > \epsilon_+ \\
&H_A : \epsilon_- \leq \tau(p') - \tau(p) \text{ and } \tau(p') - \tau(p) \leq \epsilon_+.
\end{aligned}
$$

Specifically, this joint hypothesis can be split into two one-sided hypotheses:

$$
\begin{aligned}
H_0 &: \tau(p') - \tau(p) < \epsilon_- & \qquad H_0 &: \tau(p') - \tau(p) > \epsilon_+ \\
H_A &: \tau(p') - \tau(p) \geq \epsilon_- & \qquad H_A &: \tau(p') - \tau(p) \leq \epsilon_+.
\end{aligned} \tag{14}
$$

As in my empirical applications, SEs for the IHB estimates $\tau(p') - \tau(p)$ can be obtained via (cluster) bootstrapping procedures, and the hypotheses in Equation 14 can be tested using the two one-sided tests procedure (Schuirmann 1987). If both one-sided tests of the hypotheses in Equation 14 are statistically significant at level α, then there is size-α statistically significant evidence that the IHB is small enough that the statistical significance conclusions concerning the TE will not change if experiments are hypothetically incentivized (Berger & Hsu 1996). Fitzgerald (2024) provides the `tsti` command in Stata and the `tst` command in the `eqtesting` R package to conduct such testing.[9] This approach can also be easily augmented to statistically significantly bound other hypothetical bias measures.

Further, when TEs of interest are statistically significantly different from zero under real incentive conditions, non-inferiority approaches can provide statistically significant evidence that IHBs are bounded in such a way that hypothetical incentivization will not change statistical significance conclusions. Returning to the example from the previous paragraph, presume that $\hat{\tau}(p)$ is statistically significantly greater than zero using a two-sided test. However, suppose that if the point estimate of $\hat{\tau}(p)$ were to decrease by $\epsilon > 0$, that $\hat{\tau}(p)$ would no longer be statistically significantly different from zero. One can test to ensure that the IHB is not less than $-\epsilon$ using a non-inferiority test (see Walker & Nowacki 2011). That is,

[9] To download `tsti`, see https://github.com/jack-fitzgerald/tsti, and to download `eqtesting`, see https://github.com/jack-fitzgerald/eqtesting.

one can assess the hypotheses

$$\begin{aligned} H_0 &: \tau(p') - \tau(p) < -\epsilon \\ H_A &: \tau(p') - \tau(p) \geq -\epsilon. \end{aligned} \tag{15}$$

After IHB SEs are obtained via (cluster) bootstrapping, one can assess the hypotheses in Equation 15 using a one-sided test. If the researcher finds statistically significant evidence for this one-sided test, then there is statistically significant evidence that the IHB is not negative enough for hypothetical incentivization to change the statistical significance of the TE estimate.

The key difference between the non-inferiority approach and the equivalence testing approach is that the non-inferiority approach is only concerned with bounding the IHB in one direction (see Walker & Nowacki 2011). If $\hat{\tau}(p)$ is already statistically significantly greater than zero, the non-inferiority approach presumes that the researcher does not care about positive IHBs, as such IHBs would still yield statistically significant TE estimates even if the experiment is hypothetically incentivized (provided that hypothetical incentivization does not yield considerably positive TESEB). This procedure can be naturally inverted if $\hat{\tau}(p)$ is statistically significantly less than zero, rather than greater than zero. Similar non-inferiority approaches exist for testing violations of the parallel trends assumption in difference-in-differences analyses (see Bilinski & Hatfield 2020).

6.4 How Useful Is This Research Agenda?

The usefulness of analyses of the form proposed in the previous subsection depends on such analyses' capacity to inform future experiments as to whether TE-relevant hypothetical biases threaten to change conclusions if real incentives are omitted. Part of the reason why recent hypothetical bias studies have gained traction is because their findings have been misinterpreted as being widely applicable. For example, it is cost-effective to run a hypothetical bias experiment on time preferences if finding statistically insignificant CHB for time preferences in one experiment truly means that omitting real incentives does not matter for all future experiments where (TEs on) time preferences are of interest.

However, individual findings from hypothetical bias studies are not widely portable. Evidence on TE-relevant hypothetical biases for one intervention's TE on a given outcome does not necessarily transfer to other interventions' TEs on that outcome, nor onto that intervention's TE for other outcomes. For an older study's findings on IHB and TESEB concerning outcome Y_i and intervention D_i to be portable to a newer experiment, that newer experiment must be utilizing the same Y_i and D_i in the same experimental setting. Thus unless outcome Y_i is often combined with treatment D_i, evidence concerning IHB for the TE of D_i on Y_i is not likely to be relevant for future experiments.

Even if a given treatment and outcome are often combined, sufficiently bounding TE-relevant hypothetical biases for that TE in one experiment does not necessarily imply that this same bounding will be sufficient in another experiment. For example, the portability of non-inferiority testing results from an older experiment to a newer experiment relies both on the *distance* between $(\tau(p') - \tau(p))$ and $-\epsilon$ being weakly greater in the newer experiment than in the older experiment, and on the *standard error* of $\tau(p)$ being weakly smaller in the newer experiment than in the older experiment. The former condition can meaningfully break down if incentivized TE point estimate $\tau(p)$ moves closer to significance thresholds in the newer experiment than in the older experiment, and both conditions can break down if either the TE point estimate $\tau(p)$ or the IHB point estimate $\tau(p') - \tau(p)$ is less precisely-estimated in the newer experiment than in the older experiment. Both of these events can in principle happen by chance simply due to sampling variability, even if the newer experiment is (weakly) better-powered than the older experiment to detect the same effect size. In summary, hypothetical bias boundings from non-inferiority/equivalence testing are not generally portable across experiments because even holding intervention D_i and outcome Y_i fixed, there is virtually never one single 'true' TE of D_i on Y_i, and even if there is, this 'true' TE is not likely to be observed exactly in any given experiment.

The only setting where a significantly bounded IHB reliably rules out the prospect that omitting real incentives will change experimental conclusions on a TE is *within* an experiment, but credibly obtaining this evidence is likely more expensive than just fully incentivizing the experiment from the start. Suppose that an experimental economist seeks to experimentally examine the TE of D_i on Y_i, and wishes to minimize costs subject to a power constraint. To that end, they explore the possibility of reducing costs by randomizing some participants into a fully unincentivized version of the experiment, which would permit them to show that IHBs are sufficiently bounded for their experiment.[10] The economist knows that convincing other economists that their observed TEs are externally valid will require such TEs to be observed in a fully incentivized setting. Thus to ensure that they will still obtain precise TE estimates for participants facing real incentives in the event that hypothetical bias is not sufficiently well-bounded, the economist recruits just enough participants to sufficiently power the experiment under real incentive conditions. Because they recruit just enough participants to meet the power constraint under real incentive conditions, the economist cannot spare participants from the real incentives condition while still satisfying the power constraint. Thus to run the experiment with a hypothetically-incentivized arm, the economist must recruit more participants. Even if no task-related incentives are provided for these new participants, if there are any costs whatsoever for recruiting the new participants – either in time or in money – then the economist will be strictly worse off by recruiting participants for a hypothetically-incentivized arm. In short, attempting to demonstrate that IHBs are sufficiently bounded for an experiment defeats the purpose of doing so, as under the standard constraints of experimental economics, attempting to reduce costs via hypothetically incentivizing some participants in fact incurs more costs.

7 Conclusion

This paper shows that the recent hypothetical bias literature does not justify leaving experimental choices in most modern experiments unincentivized. I provide a new taxonomy of experiments, distinguishing between ‘elicitation experiments’ where TEs are not of interest and ‘intervention experiments’ where TEs are of interest. I show econometrically and empirically that classical hypothetical bias measures can wildly misidentify TE-relevant hypothetical biases, and that traditional ways of investigating hypothetical bias are typically unproductive for informing future experimental practice.

[10] This prospect is distinct from probabilistic incentivization, where all participants know that they have some nonzero probability of being incentivized; I discuss probabilistic incentivization in further detail in Section 7.

Experimental economics' norms in favor of providing real incentives for experimental choices are still useful for ensuring that experimental TEs are externally valid. Experimental economists can often substantially reduce the costs of running experiments by completely omitting real incentives. However, the experimental economics literature is rich with examples where real incentives meaningfully impact TEs on human decision-making. For instance, Campos-Mercade et al. (2024) find that stated and revealed preferences for vaccination strongly positively correlate, but whereas the impact of donation-based incentives on stated vaccination preferences is significantly negative, the impact of the same treatment on actual vaccination behavior is significantly positive. Given that 'incentives matter' is one of the key tenets of economics, it is useful for experimental economists to presume that incentivization schemes may meaningfully impact experimental TEs, and thus to constrain experimental economists' choices by functionally requiring real incentives to be provided for experimental choices before experimental TEs are trusted. By coordinating researchers' expectations on hypothetical bias, this norm often has the added benefit of preventing researchers from pursuing the largely unproductive route of trying to empirically show that hypothetical biases for each experimental setting of interest are negligible in an effort to demonstrate that omitting real incentives is methodologically justified (see Section 6.4).

However, norms requiring full experimental incentivization have meaningful exclusionary impacts on scholars who can not secure sufficient research funding (Bardsley et al. 2009). This is a non-negligible factor in the overrepresentation of scholars and samples from Western, educated, industrialized, rich, and democratic (WEIRD) countries in the published experimental economics literature (see Henrich, Heine, & Norenzayan 2010). Given that TEs observed in WEIRD countries do not always generalize in non-WEIRD countries, this exclusionary consequence partially decreases the generalizability of TEs observed in the experimental economics literature (Henrich, Heine, & Norenzayan 2010).

One meaningful change in methodological norms that would decrease costs while still potentially preserving the external validity afforded by real incentives is disciplinary permission to use probabilistic incentivization. This involves (honestly) informing all participants that only a randomly-selected subset of their experimental choices will be incentivized, and/or that only a randomly-selected subset of the sample will receive task-related incentives for their experimental choices. This method has become more common in recent years, and has been the subject of recent methodological recommendations (see Charness, Gneezy, & Halladay 2016; Voslinsky & Azar 2021).

Probabilistic incentivization is a popular subject of empirical examination in experimental economics, but the empirical literature on probabilistic incentivization suffers from all of the same problems as the historical hypothetical bias literature. Principally, most experiments on the impacts of probabilistic incentivization vary no interventions other than incentivization schemes, and only report evidence of CHB (see Clot, Grolleau, & Ibanez 2018; Anderson et al. 2023; Umer 2023). My identification results in Section 4 demonstrate that factorial experiments that vary both the intervention(s) of interest and the incentivization scheme are necessary for identifying TE-relevant hypothetical biases arising from probabilistic incentivization. Further, Sections 6.2 and 6.3 make clear that estimates of these biases need to be tested using non-inferiority and equivalence testing approaches. However, as I discuss in Section 6.4, this line of research is not particularly productive, as hypothetical bias experiments are neither cost-effective nor credibly informative about future experiments, regardless of whether the incentivization scheme of interest is pure hypothetical incentives or probabilistic incentives.

Rather than waiting on empirical evidence on hypothetical biases in probabilistically-incentivized experiments that will be costly to obtain and will probably be uninformative anyways, it is thus likely more productive for experimental economics to simply establish an explicit norm that probabilistically-incentivized experiments are acceptable in experimental economics. This is not a significant departure from current practice, as many experimental economists already approach experiments with the implicit understanding that probabilistic incentivization yields decision frames for participants that ensure externally-valid treatment effects. For example, the seminal Holt & Laury (2002) multiple price list for risk preference elicitation employs probabilistic incentives. For participants in real incentive conditions, only one of the ten lottery choices is randomly selected to be played out for real stakes. This multiple price list is in widespread use; at time of writing, Web of Science reports over 2900 citations on Holt & Laury (2002). Thousands of TE estimates on risk aversion parameters, and thousands of other TE estimates where risk aversion parameters are controls in the model, are reliant upon the probabilistically-incentivized Holt & Laury (2002) multiple price list or other subsequent derivations thereof. Any economist confident in the generalizability of these TEs should be similarly confident in the generalizability of TE estimates arising from other probabilistically-incentivized experiments. This is a setting where norms, rather than empirics, will provide better guidance for experimental practice, accommodating incentivization schemes that strike an ideal balance between ensuring externally-valid experimental TEs and making experimental economics more accessible to scholars from all sorts of institutions around the world.

References

Alesina, Alberto, Stefanie Stantcheva, and Edoardo Teso (2018). “Intergenerational mobility and preferences for redistribution”. *American Economic Review* 108.2, pp. 521–554. DOI: 10.1257/aer.20162015.

Allais, M. (Oct. 1953). “Le comportement de l’homme rationnel devant le risque: Critique des postulats et axiomes de L’Ecole Americaine”. *Econometrica* 21.4, p. 503. DOI: 10.2307/1907921.

Altman, D. G. and J. M. Bland (1995). “Statistics notes: Absence of evidence is not evidence of absence”. *BMJ* 311.7003, pp. 485–485. DOI: 10.1136/bmj.311.7003.485.

Andel, Chantal E.E. van, Joshua M. Tybur, and Paul A.M. Van Lange (2016). “Donor registration, college major, and prosociality: Differences among students of economics, medicine and psychology”. *Personality and Individual Differences* 94, pp. 277–283. DOI: 10.1016/j.paid.2016.01.037.

Anderson, Lisa R. et al. (2023). “Pay every subject or pay only some?” *Journal of Risk and Uncertainty* 66.2, pp. 161–188. DOI: 10.1007/s11166-022-09389-6.

Andreoni, James and John Miller (Mar. 2002). “Giving according to GARP: An experimental test of the consistency of preferences for altruism”. *Econometrica* 70.2, pp. 737–753. DOI: 10.1111/1468-0262.00302.

Angrist, Joshua D and Jörn-Steffen Pischke (May 2010). “The credibility revolution in empirical economics: How better research design is taking the con out of econometrics”. *Journal of Economic Perspectives* 24.2, pp. 3–30. DOI: 10.1257/jep.24.2.3.

Arel-Bundock, Vincent, Noah Greifer, and Etienne Bacher (2024). *marginaleffects: Predictions, comparisons, slopes, marginal means, and hypothesis tests.* DOI: 10.32614/CRAN.package.marginaleffects.

Ashton, Robert H. (1990). “Pressure and performance in accounting decision settings: Paradoxical effects of incentives, feedback, and justification”. *Journal of Accounting Research* 28, pp. 148–180. DOI: 10.2307/2491253.

Bardsley, Nicholas et al. (2009). “Incentives in experiments”. *Experimental economics: Rethinking the rules.* 1st ed. Princetown University Press, pp. 244–285.

Becker, Gordon M., Morris H. DeGroot, and Jacob Marschak (1964). “Measuring utility by a single-response sequential method”. *Behavioral Science* 9.3, pp. 226–232. DOI: 10.1002/bs.3830090304.

Berger, Roger L. and Jason C. Hsu (1996). “Bioequivalence trials, intersection-union tests and equivalence confidence sets”. *Statistical Science* 11.4. DOI: 10.1214/ss/1032280304.

Bilinski, Alyssa and Laura A Hatfield (2020). “Nothing to see here? Non-inferiority approaches to parallel trends and other model assumptions”. *arXiv*. DOI: 10.48550/arXiv.1805.03273.

Brañas-Garza, Pablo et al. (2022). “Paid and hypothetical time preferences are the same: Lab, field and online evidence”. *Experimental Economics* 26.2, pp. 412–434. DOI: 10.1007/s10683-022-09776-5.

Cadena, Brian C. and Benjamin J. Keys (2015). “Human capital and the lifetime costs of impatience”. *American Economic Journal: Economic Policy* 7.3, pp. 126–153. DOI: 10.1257/pol.20130081.

Camerer, Colin F. and Robin 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, pp. 7–42. DOI: 10.1023/a:1007850605129.

Campos-Mercade, Pol et al. (2024). “Incentives to vaccinate”. *NBER Working Paper Series* 32899. DOI: 10.3386/w32899.

Ceccato, Smarandita et al. (2018). “Social preferences under chronic stress”. *PLOS ONE* 13.7, e0199528. DOI: 10.1371/journal.pone.0199528.

Chamberlin, Edward H. (Apr. 1948). “An experimental imperfect market”. *Journal of Political Economy* 56.2, pp. 95–108. DOI: 10.1086/256654.

Charness, Gary, Uri Gneezy, and Brianna Halladay (2016). “Experimental methods: Pay one or pay all”. *Journal of Economic Behavior Organization* 131, pp. 141–150. DOI: 10.1016/j.jebo.2016.08.010.

Charness, Gary and Mark Pingle (2022). *The Art of Experimental Economics: Twenty top papers reviewed*. Routledge.

Clot, Sophie, Gilles Grolleau, and Lisette Ibanez (2018). "Shall we pay all? An experimental test of random incentivized systems". *Journal of Behavioral and Experimental Economics* 73, pp. 93–98. DOI: 10.1016/j.socec.2018.01.004.

Cummings, Ronald G. et al. (1997). "Are hypothetical referenda incentive compatible?" *Journal of Political Economy* 105.3, pp. 609–621. DOI: 10.1086/262084.

Edwards, Ward (1953). "Probability-preferences in gambling". *The American Journal of Psychology* 66.3, pp. 349–364. DOI: 10.2307/1418231.

Enke, Benjamin et al. (2021). *Replication data for: Cognitive biases: Mistakes or missing stakes?* Dataset V1. Cambridge, MA, U.S.A.: Harvard Dataverse. DOI: 10.7910/DVN/HBQLA6.

— (2023). "Cognitive biases: Mistakes or missing stakes?" *Review of Economics and Statistics* 105 (4), pp. 818–832. DOI: 10.1162/rest_a_01093.

Fang, Di et al. (2020). "On the use of virtual reality in mitigating hypothetical bias in choice experiments". *American Journal of Agricultural Economics* 103.1, pp. 142–161. DOI: 10.1111/ajae.12118.

Fitzgerald, Jack (May 2024). *The need for equivalence testing in economics.* Discussion Paper 125. Institute for Replication. URL: https://www.econstor.eu/handle/10419/296190.

Flood, Merrill M. (Oct. 1958). "Some experimental games". *Management Science* 5.1, pp. 5–26. DOI: 10.1287/mnsc.5.1.5.

Forsythe, Robert et al. (1994). "Fairness in simple bargaining experiments". *Games and Economic Behavior* 6.3, pp. 347–369. DOI: 10.1006/game.1994.1021.

Gelman, Andrew and Hal Stern (2006). "The difference between "significant" and "not significant" is not itself statistically significant". *The American Statistician* 60.4, pp. 328–331. DOI: 10.1198/000313006x152649.

Golsteyn, Bart H.H., Hans Grönqvist, and Lena Lindahl (2014). "Adolescent time preferences predict lifetime outcomes". *The Economic Journal* 124.580, F739–F761. DOI: 10.1111/ecoj.12095.

Guala, Francesco (2001). "Clear-cut designs versus the uniformity of experimental practice". *Behavioral and Brain Sciences* 24.3, pp. 412–413. DOI: 10.1017/s0140525x01334143.

Hackethal, Andreas et al. (2023). “On the role of monetary incentives in risk preference elicitation experiments”. *Journal of Risk and Uncertainty* 66.2, pp. 189–213. DOI: 10.1007/s11166-022-09377-w.

Harrison, Glenn W et al. (2005). “Risk aversion and incentive effects: Comment”. *American Economic Review* 95.3, pp. 897–901. DOI: 10.1257/0002828054201378.

Hausman, Jerry (2012). “Contingent valuation: From dubious to hopeless”. *Journal of Economic Perspectives* 26.4, pp. 43–56. DOI: 10.1257/jep.26.4.43.

Hayes, Andrew F. and Li Cai (2007). “Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation”. *Behavior Research Methods* 39.4, pp. 709–722. DOI: 10.3758/bf03192961.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan (2010). “The weirdest people in the world?” *Behavioral and Brain Sciences* 33.2–3, pp. 61–83. DOI: 10.1017/s0140525x0999152x.

Hertwig, Ralph and Andreas Ortmann (2001). “Experimental practices in economics: A methodological challenge for psychologists?” *Behavioral and Brain Sciences* 24.3, pp. 383–403. DOI: 10.1017/s0140525x01004149.

Holt, Charles A and Susan K Laury (2002). “Risk aversion and incentive effects”. *American Economic Review* 92.5, pp. 1644–1655. DOI: 10.1257/000282802762024700.

Irwin, Julie R., Gary H. McClelland, and William D. Schulze (1992). “Hypothetical and real consequences in experimental auctions for insurance against low-probability risks”. *Journal of Behavioral Decision Making* 5.2, pp. 107–116. DOI: 10.1002/bdm.3960050203.

Jamal, Karim and Shyam Sunder (1991). “Money vs gaming: Effects of salient monetary payments in double oral auctions”. *Organizational Behavior and Human Decision Processes* 49.1, pp. 151–166. DOI: 10.1016/0749-5978(91)90046-v.

Kuziemko, Ilyana et al. (2015). “How elastic are preferences for redistribution? Evidence from randomized survey experiments”. *American Economic Review* 105.4, pp. 1478–1508. DOI: 10.1257/aer.20130360.

Li, Lunzheng, Zacharias Maniadis, and Constantine Sedikides (2021). “Anchoring in economics: A meta-analysis of studies on willingness-to-pay and willingness-to-accept”. *Journal of Behavioral and Experimental Economics* 90, p. 101629. DOI: 10.1016/j.socec.2020.101629.

List, John A (2001). "Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for Sportscards". *American Economic Review* 91.5, pp. 1498–1507. DOI: 10.1257/aer.91.5.1498.

MacKinnon, James G and Halbert White (1985). "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties". *Journal of Econometrics* 29.3, pp. 305–325. DOI: 10.1016/0304-4076(85)90158-7.

Mosteller, Frederick and Philip Nogee (Oct. 1951). "An experimental measurement of utility". *Journal of Political Economy* 59.5, pp. 371–404. DOI: 10.1086/257106.

Muralidharan, Karthik, Mauricio Romero, and Kaspar Wüthrich (2023). "Factorial designs, model selection, and (incorrect) inference in randomized experiments". *The Review of Economics and Statistics*, pp. 1–44. DOI: 10.1162/rest_a_01317.

Murphy, James J. et al. (2005). "A meta-analysis of hypothetical bias in stated preference valuation". *Environmental & Resource Economics* 30.3, pp. 313–325. DOI: 10.1007/s10640-004-3332-z.

Roth, Alvin E (1995). "Introduction to Experimental Economics". *Handbook of Experimental Economics*. Princeton University Press, pp. 3–109.

Rousseas, Stephen W. and Albert G. Hart (Aug. 1951). "Experimental verification of a composite indifference map". *Journal of Political Economy* 59.4, pp. 288–318. DOI: 10.1086/257092.

Schuirmann, Donald J. (1987). "A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability". *Journal of Pharmacokinetics and Biopharmaceutics* 15.6, pp. 657–680. DOI: 10.1007/bf01068419.

Schwieren, Christiane et al. (2018). *Social preferences under chronic stress*. Dataset V1. Heidelberg, Germany: heiDATA. DOI: 10.11588/data/F68JZT.

Scott, W.E, Jiing-Lih Farh, and Philip M Podsakoff (1988). "The effects of "intrinsic" and "extrinsic" reinforcement contingencies on task behavior". *Organizational Behavior and Human Decision Processes* 41.3, pp. 405–425. DOI: 10.1016/0749-5978(88)90037-4.

Sefton, Martin (1992). "Incentives in simple bargaining games". *Journal of Economic Psychology* 13.2, pp. 263–276. DOI: 10.1016/0167-4870(92)90033-4.

Smith, Vernon L (Apr. 1962). "An experimental study of competitive market behavior". *Journal of Political Economy* 70.2, pp. 111–137. DOI: 10.1086/258609.

— (1976). "Experimental economics: Induced value theory". *American Economic Review* 66.2, pp. 274–279.

— (Dec. 1982). "Microeconomic systems as an experimental science". *American Economic Review* 72.5, pp. 923–955.

— (1965). "Experimental auction markets and the Walrasian hypothesis". *Journal of Political Economy* 73.4, pp. 387–393. DOI: 10.1086/259041.

Smith, Vernon L. and James M. Walker (1993). "Monetary rewards and decision cost in experimental economics". *Economic Inquiry* 31.2, pp. 245–261. DOI: 10.1111/j.1465-7295.1993.tb00881.x.

Sunde, Uwe et al. (2022). "Patience and comparative development". *The Review of Economic Studies* 89.5, pp. 2806–2840. DOI: 10.1093/restud/rdab084.

Svorenčík, Andrej and Harro Maas (2016). *The making of experimental economics: Witness seminar on the emergence of a field*. Springer.

Thurstone, L. L. (May 1931). "The indifference function". *The Journal of Social Psychology* 2.2, pp. 139–167. DOI: 10.1080/00224545.1931.9918964.

Umer, Hamza (2023). "Effectiveness of random payment in experiments: A meta-analysis of dictator games". *Journal of Economic Psychology* 96, p. 102608. DOI: 10.1016/j.joep.2023.102608.

Van Lange, Paul A.M., Michaéla Schippers, and Daniel Balliet (2011). "Who volunteers in psychology experiments? An empirical review of prosocial motivation in volunteering". *Personality and Individual Differences* 51.3, pp. 279–284. DOI: 10.1016/j.paid.2010.05.038.

Voslinsky, Alisa and Ofer H. Azar (2021). "Incentives in experimental economics". *Journal of Behavioral and Experimental Economics* 93, p. 101706. DOI: 10.1016/j.socec.2021.101706.

Walker, Esteban and Amy S. Nowacki (2011). "Understanding equivalence and noninferiority testing". *Journal of General Internal Medicine* 26.2, pp. 192–196. DOI: 10.1007/s11606-010-1513-8.

Wasserstein, Ronald L. and Nicole A. Lazar (2016). “The ASA statement on p-values: Context, process, and purpose”. *The American Statistician* 70.2, pp. 129–133. DOI: 10.1080/00031305.2016.1154108.

Wright, William F and Urton Anderson (1989). “Effects of situation familiarity and financial incentives on use of the anchoring and adjustment heuristic for probability assessment”. *Organizational Behavior and Human Decision Processes* 44.1, pp. 68–82. DOI: 10.1016/0749-5978(89)90035-6.