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

Jack Fitzgerald*

October 2, 2024

Abstract

Recent studies showing that some outcome variables do not statistically significantly differ between real-stakes and hypothetical-stakes conditions have raised methodological challenges to experimental economics' disciplinary norm that experimental choices should be incentivized with real stakes. I show that the hypothetical bias measures estimated in these studies do not econometrically identify the hypothetical biases that matter in most modern experiments. Specifically, traditional 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, traditional hypothetical bias measures are uninformative; real stakes matter if and only if TEs differ between stakes conditions. I demonstrate that traditional hypothetical bias metrics are often 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 stakes conditions does not imply that all TEs on that outcome are unaffected by hypothetical stakes. Therefore, the recent hypothetical bias literature does not justify abandoning real stakes in most modern experiments. Maintaining norms that favor completely or probabilistically providing real stakes for experimental choices is 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, Andreas Ortmann, 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

Incentivizing experimental choices with real stakes is a key feature of experimental economics. This approach is a long-standing norm in experimental economics, as participants' desire to optimize real-world outcomes can improve the generalizability of experimental behavior by overpowering biases known to emerge in experimental environments (see Smith 1976; Smith 1982; Roth 1995; Camerer & Hogarth 1999; Hertwig & Ortmann 2001; Schram 2005; Bardsley et al. 2009; Charness, Gneezy, & Halladay 2016; Svorenčík & Maas 2016; Clot, Grolleau, & Ibanez 2018). However, this norm is starting to shift. Top economics publications are becoming increasingly open to publishing results from hypothetical-stakes experiments (e.g., see Golsteyn, Grönqvist, & Lindahl 2014; Cadena & Keys 2015; Kuziekmo et al. 2015; Alesina, Stantcheva, & Teso 2018; Sunde et al. 2022). Recent research also shows that some outcome variables do not statistically significantly differ on average between real-stakes and hypothetical-stakes conditions (see Matousek, Havranek, & Irsova 2022; Brañas-Garza et al. 2023; Enke et al. 2023; Hackethal et al. 2023). Citing some of this recent hypothetical bias research (in particular Matousek, Havranek, & Irsova 2022), the announcement for *Experimental Economics*' special issue on incentivization states: "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 existing hypothetical bias literature does not statistically support omitting real stakes in most modern experiments. I begin by distinguishing two types of experiments. In 'elicitation experiments', no intervention is varied, and treatment effects (TEs) are not of interest. In contrast, 'intervention experiments' vary at least one intervention with the goal of measuring its TE. Elicitation experiments dominated early experimental economics research, and though they remain important to this day, most modern economic experiments are intervention experiments.

Econometrically, traditional tests for hypothetical bias do not identify the hypothetical biases that matter for intervention experiments. I show that the hypothetical bias relevant for intervention experiments is the interaction effect between hypothetical stakes and the treat-

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

ment of interest. However, most traditional hypothetical bias experiments cannot identify this interaction effect. Typically, these experiments randomize participants into either real-stakes or hypothetical-stakes conditions, elicit an outcome, and test whether the difference in average outcomes between the two conditions is statistically significant (e.g., Brañas-Garza et al. 2023; Hackethal et al. 2023). I show that this difference is the average marginal effect of hypothetical stakes on the outcome, which has no general relationship with the interaction effect between hypothetical stakes and any treatment of interest. This makes sense for two reasons. First, a researcher cannot identify an interaction effect if all the researcher knows is the average marginal effect of one of the two variables in the interaction. Second, it is unrealistic to expect hypothetical stakes to affect every possible intervention's TE on a given outcome in the exact same way.

Empirically, TE-irrelevant hypothetical bias measures often meaningfully misidentify TE-relevant hypothetical biases. I reanalyze data from three recent hypothetical bias experiments that vary both a treatment of interest and hypothetical stakes. These experiments allow me to directly estimate the interaction effects between hypothetical stakes and treatments of interest, and to compare these interaction effects with the TE-irrelevant hypothetical bias estimates typically produced in hypothetical bias experiments. I find that TE-irrelevant hypothetical bias measures often yield different conclusions than TE-relevant hypothetical bias measures. In some cases, TE-irrelevant hypothetical bias measures even exhibit sign flips when compared to TE-relevant hypothetical bias measures. That is, TE-irrelevant hypothetical bias estimates are sometimes positive even when TE-relevant hypothetical biases are negative (and vice versa).

These findings raise doubts about the practical value of recent advances in the hypothetical bias literature. My econometric results show that recent studies finding no statistically significant differences in certain outcomes between real-stakes and hypothetical-stakes conditions do not justify the broader conclusion that real stakes 'do not matter' for all TEs on those outcomes. Researchers who abandon real experimental stakes in their intervention experiments based on these findings may be misled, and TEs estimated in these experiments may be confounded by meaningful hypothetical biases. 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, it is also unproductive and uninformative to conduct hypothetical bias experiments with the goal of 'paving the way' for future researchers to abandon real stakes in their experiments. Therefore, it remains useful to maintain existing norms in experimental economics that favor incentivizing experimental choices with real stakes. Because incentivizing all experimental choices for all participants is too expensive for some researchers, it is also likely beneficial to augment these norms by allowing researchers to incentivize experimental choices with real stakes probabilistically.

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. Section 4 establishes econometrically why these traditional methods for measuring hypothetical bias fail to identify TE-relevant hypothetical biases. Section 5 provides three empirical applications demonstrating that TE-relevant and TE-irrelevant hypothetical bias measures often differ. Section 6 discusses the implications of my findings for the future of hypothetical bias research. Section 7 concludes with a discussion on norms and best practices for future experimental economics research.

2 Terminology and Notation

I start by establishing a simple taxonomy of experiments. 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 this paper, a 'real stakes' condition is one in which participants' experimental choices map onto real-world payoffs or consequences. In contrast, 'hypothetical stakes' conditions do not link experimental choices to real-world consequences.

I distinguish between two types of experiments, the first of which is an 'elicitation experiment.' This sort of experiment does not apply any intervention, and there are no TEs to estimate. The primary aim of an elicitation experiment is to use experimental procedures to obtain 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 the Becker, DeGroot, & Marschak (1964) procedure to obtain an incentive-compatible measure of participants' willingness to pay. This is undoubtedly an experiment, but there is no TE to speak of; 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.' Unlike 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 previous example, suppose that the researcher wants to know the effect of a specific product characteristic on willingness to pay. They could repeat the same Becker-DeGroot-Marschak experiment, but randomly assign half of the participants to consider a product with that characteristic. The researcher can then estimate the TE of that characteristic on willingness to pay by taking the difference in average willingness to pay between the two halves of the sample. This would be an intervention experiment.[2]

In general, 'hypothetical bias' can be defined as the difference in the statistic of interest resulting from a change in stakes condition S_i, which is parameterized here as a dummy variable indicating that participant i faces real stakes with probability p' instead of probability p. That is, for $p, p' \in [0, 1]$ with $p \neq p'$, I define

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

Typically, $p = 1$ and $p' = 0$, meaning $S_i = 1$ indicates pure hypothetical stakes whereas $S_i = 0$ indicates pure real stakes. I use this definition of S_i throughout the remainder of this paper for simplicity. However, this framework can be extended to examine potential biases arising from switching between any pair of probabilities that stakes are real. Because of this generalizability, the statistical framework that I introduce throughout this paper can also be used to explore hypothetical biases arising from probabilistic incentivization. I return to this point in Section 6.5. The specific bias induced by switching between stakes conditions

[2]The researcher may still be interested in descriptive statistics about Y_i in an intervention experiment. For instance, the experiment described in this paragraph is still an intervention experiment even if the researcher also wants to know the mean willingness to pay for products both with and without the characteristic of interest. So long as the experiment employs an intervention whose TE is of interest to the researcher, it is an intervention experiment.

depends on the statistic of interest.

3 Historical Measurement of Hypothetical Bias

Many early seminal contributions in experimental economics are elicitation experiments. A preponderance of economic experiments published prior to 1960 focus heavily on testing the predictions of prevailing economic theories and documenting empirical regularities observed in laboratory experiments (Roth 1995). This was largely done using elicitation experiments to measure various 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 is not to say that no intervention experiments were conducted in experimental economics' early years, but elicitation experiments certainly played a leading role.

This historical context is important because the preponderance of elicitation experiments in experimental economics' early years influenced the statistical parameters that experimental economists were interested in when disciplinary norms on experimental stakes first emerged. The influential 'Wallis-Friedman critique' of hypothetical choice menus was already published in 1942, and played a key role in prompting leading experimental economists to incentivize their experiments with real stakes (see Wallis & Friedman 1942; Svorenčík & Maas 2016; Ortmann 2016). As a result, by the end of the 1950s, experimental economists were already predominantly incentivizing their experiments with real stakes (Roth 1995). The fact that experimental economists at this time 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 stakes, as well as the ways in which they measured bias when real stakes were not provided.

Two key rationales for incentivizing experiments with real stakes emerged from this early literature. First, experimental economists believe that hypothetical stakes may affect the average preference or behavior elicited from a sample. This implies that hypothetical stakes bias the expected value of Y_i. I refer to this bias as '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 CHB $= \delta$.

CHB is a well-documented factor in economic experiments. Camerer & Hogarth (1999) provide systematic evidence of CHB, reviewing 36 studies that compare a hypothetical-stakes condition with a real-stakes control.[3] 26 of these studies (72%) show that hypothetical stakes affect the central tendency of at least one outcome. Similarly, Harrison & Ruström (2008) review 35 studies measuring CHB in experiments on willingness to pay. Only two of these studies (5.7%) report zero CHB, and 16 studies (45.7%) report statistically significantly CHB. Smith & Walker (1993) provide similar systematic evidence.

Significant CHB is found in a variety of experimental settings. These include 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 experiments. Experimental participants routinely overstate their 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 studies find that experimental outcomes do not statistically significantly differ between hypothetical-stakes and real-stakes conditions (Matousek, Havranek, & Irsova 2022; Brañas-Garza et al. 2023; Enke et al. 2023; Hackethal et al. 2023), a large body of literature demonstrates substantial risks of CHB in many experimental contexts.

The second rationale for incentivizing experiments with real stakes is reducing noise.

[3] This is a subset of the 74 experiments reviewed by Camerer & Hogarth (1999), specifically focusing on studies with a '0 vs. L' treatment, or a '0' treatment with some real stakes control. My list excludes Scott, Farh, & Podsakoff (1988) because participants were unaware of the real stakes until after the experiment concluded (Camerer & Hogarth 1999).

Experimental economists believe that participants motivated by real stakes make more careful and deliberative choices than participants facing hypothetical stakes, and thus that real stakes reduce noise in experimental outcomes (see Bardsley et al. 2009). Camerer & Hogarth (1999) note nine experiments where hypothetical stakes change the variance or convergence of experimental outcomes (usually by increasing variance or decreasing convergence). Hertwig & Ortmann (2001) identify two additional experiments where similar effects are observed. Smith & Walker (1993) survey 31 hypothetical bias studies and find that in virtually all, the variance of outcomes around theory-predicted values decreases when stakes are real.

However, the measurement of these 'noise reduction' effects is not systematic and greatly differs between studies. Some studies focus on changes in the standard deviation (SD) or variance of outcomes between stakes conditions (e.g., Wright & Anderson 1989; Ashton 1990; Irwin, McClelland, & Schulze 1992; Forsythe et al. 1994). Others assess noise by examining deviations from some theory-predicted value, such as price deviations from a competitive market price (see Edwards 1953; Smith 1962; Smith 1965; Jamal & Sunder 1991; Smith & Walker 1993). Furthermore, changes in variance between stakes conditions are typically not accompanied by a precision measure, such as a standard error (SE), to qualify the magnitude of these between-condition variance shifts.[4] It is thus unclear whether observed differences in outcome variances between stakes conditions reflect genuine effects or are simply artefacts of sampling variation.

For the purposes of this paper, I parameterize the effect of hypothetical stakes 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 stakes conditions. The SE of this estimate can be obtained via bootstrap (see Section 5 for examples). I define noise in this way because not all experimental outcomes have clear values that theoretically 'should' be observed in experimental data, whereas SDs

[4]Recent attempts to qualify the significance of differences in variance between stakes conditions often take 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.

can be used to measure noise across experimental contexts.

Most studies on the effects of real stakes in experiments focus exclusively on CHB and OSDB. Matousek, Havranek, & Irsova (2022) find that the meta-analytic average individual discount rate does not statistically significantly differ between real-stakes and hypothetical-stakes experiments. Brañas-Garza et al. (2023) find that the means and SDs of time discounting factors are not statistically significantly different between real-stakes and hypothetical-stakes conditions, and 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 no statistically significant differences in the number of correct answers on the cognitive reflection test, a base rate neglect test, or a contingent reasoning test between real-stakes and hypothetical-stakes conditions. These studies are reporting estimates of CHB, with Brañas-Garza et al. (2023) and Hackethal et al. (2023) also reporting evidence on OSDB.

Though CHB and OSDB are fully informative measures of hypothetical bias in elicitation experiments – which played early leading roles in experimental economics when norms on real stakes first emerged – most modern work in experimental economics (and experimental social sciences more broadly) is not limited to elicitation experiments. Although elicitation experiments remain important today, many researchers are now more focused on obtaining clean causal TEs from experiments than they are in simply obtaining descriptive statistics. 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 for experimental TEs.

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 while completing ignoring intervention D_i. Any statistical framework used to identify the effect of real stakes on TEs must incorporate D_i, and must allow the possibility that stakes condition S_i can influence TEs.

My econometric framework for modeling the impact of hypothetical stakes on TEs considers a simple 2x2 factorial experiment where both treatment D_i and stakes condition S_i are randomized with equal probability across participants. Following Guala (2001), I model the effects of D_i and S_i using 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: $\mu_i \perp \{D_i, S_i\}$. Participant i's treatment effect τ_i can thus be modeled in the following potential outcomes framework (see Rubin 1974; Rubin 2005):

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

Here $Y_i(D,S)$ represents the potential outcome of Y_i depending on intervention status $D \in \{0,1\}$ and stakes condition $S \in \{0,1\}$. For what follows, suppose that the statistic of interest is the average TE $\tau \equiv \mathbb{E}[\tau_i]$.

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

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

This implies that hypothetical stakes bias the TE's point estimate if and only if $\beta_3 \neq 0$. This yields an intuitive conclusion: in a factorial experiment that randomizes both an intervention and hypothetical stakes, any hypothetical bias in the point estimate of the intervention's TE is fully captured by the interaction effect between the intervention and hypothetical stakes.

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:

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

If both D_i and S_i are randomized, the CHB estimated from a simple regression of the form in Equation 3 will return $\hat{\delta} = \mathbb{E}\left[\delta_i\right]$. By Equation 9, CHB estimates can thus be decomposed in the following way:

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

CHB thus does not identify IHB, and trying to infer IHB from CHB can yield misleading conclusions. By Equation 10, if $|\beta_2|$ is large and $\beta_3 = 0$, then CHB will be large even though IHB is zero. Likewise, if $\beta_2 = -\mathbb{E}\left[D_i\right]\beta_3$, then CHB will be zero even if IHB is arbitrarily large.

In fact, CHB and IHB almost always differ. Under Equations 8 and 10, CHB and IHB differ whenever $\beta_3 \neq \mathbb{E}\left[\delta_i\right]$, which holds whenever $\beta_3 \neq \frac{\beta_2}{1-\mathbb{E}[D_i]}$. This condition holds almost always, as the interaction effect between an intervention and some moderator is virtually never exactly the same as 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 Matousek, Havranek, & Irsova (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 does not imply that hypothetical stakes induce zero bias for any intervention TEs on these outcomes. Identifying IHB requires a factorial experiment that varies both intervention D_i and stakes condition S_i in a way that permits unconfounded estimation of these treatments' individual and joint effects on Y_i. Therefore, IHB cannot be identified in experiments that only vary S_i. Further, for a given outcome variable, there is no 'one true' IHB for all interventions, 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 hypothetical bias on TE point estimates. I parameterize hypothetical bias on TE precision as '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 stakes conditions. SEs for TESEB point estimates can be estimated via bootstrapping (see Section 5 for examples).

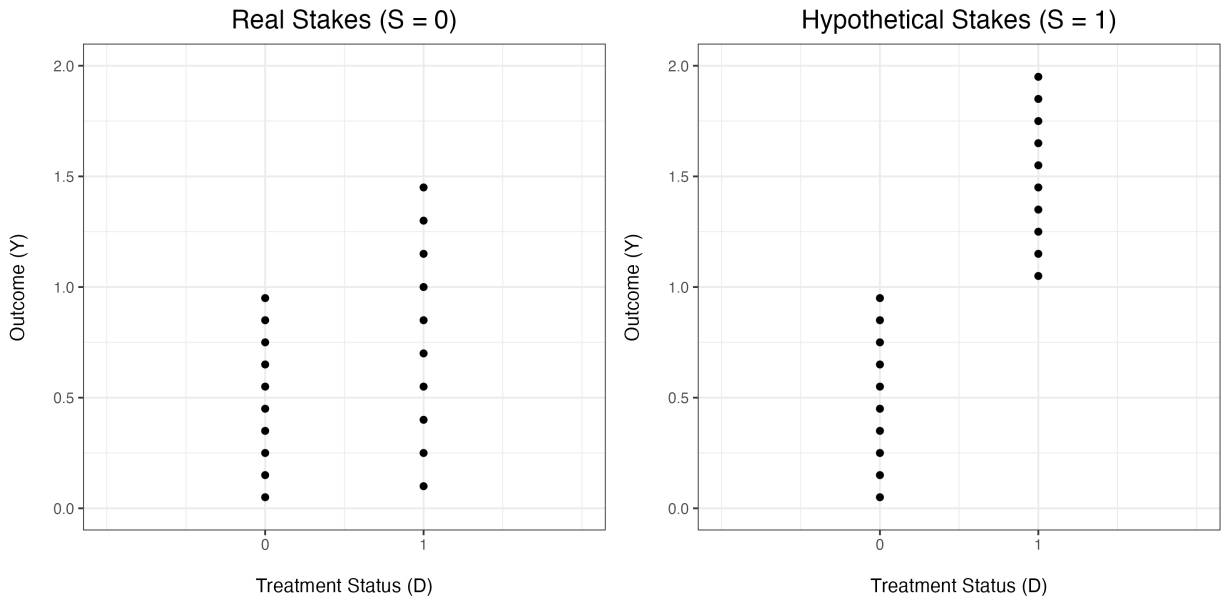
OSDB does not identify hypothetical biases on TE precision. The best way to show this is through a simple counterexample where OSDB and TESEB have opposite signs. Figure 1 displays data points from two simulated datasets, each of which contain 20 observations. In both datasets, the simulated intervention is assigned such that $D_i = 0$ for $i \in \{1, 2, \cdots 10\}$ and $D_i = 1$ for $i \in \{11, 12, \cdots 20\}$. The first dataset arises from the data-generating process

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

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 represent two halves of an experimental dataset, where D_i and S_i are randomized, where the first half (generated by the process in Equation 12) belongs to a real-stakes sample (i.e., $S_i = 0$), whereas the second half (generated by the process in Equation 13) belongs to a hypothetical-stakes sample (i.e., $S_i = 1$). It is clearly visible from Figure 1 that the outcome SD for the hypothetical-stakes sample (0.592) is higher than that in the real-stakes sample (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 hypothetical-stakes sample (0.135) than in the real-stakes sample (0.173), so TESEB



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, whereas 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

is negative.[5] This example demonstrates that OSDB does not identify TESEB, and that interpreting OSDB estimates as evidence of how hypothetical stakes affect 'noise' in TE estimates can yield misleading conclusions.

4.3 Meta-Analytic Approaches

One approach that hypothetical bias researchers use to directly estimate IHB is by meta-analytically comparing TEs from studies with and without real stakes. For instance, Li, Maniadis, & Sedikides (2021) conduct a meta-analysis of studies investigating anchoring effects on willingness to pay/accept. They find no statistically significant differences between the anchoring effects observed in studies with and without real stakes, and therefore conclude that real stakes have no discernible impact on anchoring effects. A similar approach could be used to estimate TESEBs by comparing meta-analytic averages of TE SEs under different stakes conditions, though Li, Maniadis, & Sedikides (2021) do not make this comparison.

Meta-analyses like this do not provide clean causal estimates of the impact of real stakes,

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

as the choice to incentivize an experiment with real stakes is endogenous. The identification of IHB as a simple interaction effect between treatment D_i and stakes condition S_i in Equation 8 relies on a joint unconfoundedness assumption over both the treatment the stakes condition, $\mu_i \perp \{D_i, S_i\}$. This is readily achieved *within* a factorial experiment when both the intervention and hypothetical stakes are randomly assigned. However, this unconfoundedness condition is not generally satisfied when comparing TEs *across* experiments, as experimental stakes conditions are typically not randomly assigned, and are likely correlated with other factors that simultaneously influence TEs and their SEs.

One important factor that likely confounds meta-analytic IHB estimates is academic disciplines. Naturally, some disciplines are more likely to provide real experimental stakes than others, and these disciplines meaningfully differ on various important dimensions, including participant pools and procedural norms in experimentation (see Hertwig & Ortmann 2001; Bardsley et al. 2009). To fix a simple example, consider a meta-analytic dataset where all experiments employing real stakes are run by economists, whereas all experiments employing hypothetical stakes are run by psychologists. Further, suppose that the economics experiments recruit economics students, whereas the psychology experiments recruit psychology students. In order to credibly interpret the difference in TEs between these groups of experiments as a causal effect of hypothetical stakes, 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 assumption is untenable; economics students differ from psychology students in important ways, and the same treatment can affect economics students and psychology students in significantly different ways (Van Lange, Schippers, & Balliet 2011; van Andel, Tybur, & Van Lange 2016). Therefore, meta-analytic differences between TEs thus do not generally provide clean identification of IHBs. For similar reasons, meta-analytic differences between TE SEs do not generally provide clean identification of TESEBs.

5 Empirical Applications

It is challenging to find experiments where clean estimates of IHB and TESEB can be respectively compared with CHB and OSDB. The vast majority of hypothetical bias experiments only vary stakes conditions and introduce no other interventions (e.g., see Walker & Smith 1993; Camerer & Hogarth 1999; Hertwig & Ortmann 2001; Harrison & Rutström 2008; Brañas-Garza et al. 2023; Hackethal et al. 2023). This makes it impossible to obtain IHB or TESEB estimates in these studies (see Sections 4.1 and 4.2). Even when experiments vary both an intervention and hypothetical stakes, the interaction effects between these treatments are almost never reported. This is likely a consequence of publication bias against null results (see Fanelli 2012; Franco, Malhotra, & Simonovits 2014; Andrews & Kasy 2019; Chopra et al. 2024). For example, IHBs are interaction effects, which are notoriously noisy and difficult to sufficiently power (Muralidharan, Romero, & Wüthrich 2023). Many IHB estimates are thus likely statistically insignificant, meaning that many likely go unreported. Estimating IHB and TESEB in suitable studies that do not report estimates of these biases requires replication data. However, virtually no published articles provide full data/code unless their journal mandates data-sharing, many data-sharing policies are fairly recent, and a considerable number of economics journals do not mandate data-sharing (Askarov et al. 2023; Brodeur, Cook, & Neisser 2024).

To assess whether TE-irrelevant hypothetical bias measures misidentify TE-relevant hypothetical biases in practice, I replicate three of the first hypothetical bias experiments I can find which permit me to directly estimate CHB, IHB, OSDB, and TESEB. These studies have publicly available replication data and leverage factorial designs that simultaneously manipulate both hypothetical stakes and another intervention. 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 stakes are hypothetical or into a room where stakes are real, after which partic-

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	-9.547	2.312	-2.39	348
			(2.575)	(5.16)	(1.55)	(1.423)	
Fang et al. (2021)	Purchasing yogurt (0/1)	Virtual reality (0/1)	0.182	0.049	-0.162	0.579	1024
			(0.04)	(0.08)	(0.034)	(1.439)	
Enke et al. (2023)	Answer (0-100)	Numerical anchor (0-100)	6.04	0.019	0.985	0.154	626
			(2.338)	(0.074)	(1.356)	(2.052)	

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

ipants 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 stakes, 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 $\%\text{Trans}_i$ 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 stakes 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 stakes are hypothetical, but the SE of the TE of 'give' framing on endowment transfers is smaller when stakes 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 hypothetical-stakes 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-stakes condition whereas the other is a real-stakes condition. In the fourth and fifth conditions, participants make product decisions in a virtual reality supermarket setting; stakes are real in one of these two virtual reality conditions whereas stakes are hypothetical in the other. 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; Harrison & Rutström 2008; 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 stakes 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 condition where stakes are real 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.[6] However, Enke et al. (2023) also examine the impact of stakes 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;[7] one of these questions is faced under

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

[7] Questions include "Is the time (in minutes) it takes for light to travel from the Sun to the planet Jupiter

real-stakes conditions and the other is faced under hypothetical-stakes 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 anchor. Participants must thereafter provide an exact numerical answer to the question. In the real-stakes 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 stakes, I exclude the portion of the sample subjected to the high-stakes 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-

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).

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 hypothetical stakes 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 real-stakes to hypothetical-stakes 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 stakes 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 with real stakes. For example, Matousek, Havranek, & Irsova (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 with real stakes, reasoning that the statistically insignificant CHB estimates in Matousek, Havranek, & Irsova (2022) and Brañas-Garza et al. (2023) are evidence that real stakes '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 stakes 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 hypothetical-stakes 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 stakes have 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. 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 stakes? What most researchers ultimately care about is whether the conclusions that they make about TEs are meaningfully impacted by stakes conditions. In practice, this means that before a researcher chooses to omit real stakes, 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 stakes.

When TEs of interest are not statistically significantly different from zero under real stakes 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 stakes condition S_i are exogenously varied amongst participants i. Further, suppose that the TE estimate of interest under real stakes 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:

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

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 have no real stakes (Berger & Hsu 1996). Fitzgerald (2024) provides the `tsti` command in Stata and the `tst` command in the `eqtesting` R package to conduct such testing.[8] 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-stakes conditions, non-inferiority approaches can provide statistically significant evidence that IHBs are bounded in such a way that hypothetical stakes 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, one

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

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 stakes 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 not incentivized with real stakes (provided that hypothetical stakes do 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 stakes 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 stakes 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 real-stakes 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 stakes 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 hypothetical-stakes version of the experiment, which would permit them to show that IHBs are sufficiently bounded for their experiment.[9] The economist knows that convincing other economists that their observed TEs are externally valid will require such TEs to be observed in a real-stakes setting. Thus to ensure that they will still obtain precise TE estimates for participants facing real stakes in the event that hypothetical bias is not sufficiently well-bounded, the economist recruits just enough participants to sufficiently power the experiment under real-stakes conditions. Because they recruit just enough participants to meet the power constraint under real-stakes conditions, the economist cannot spare participants from the real-stakes condition while still satisfying the power constraint. Thus to run the experiment with a hypothetical-stakes arm, the economist must recruit more participants. Even if no real stakes 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 hypothetical-stakes 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 omitting real stakes for some participants in fact incurs more costs.

6.5 Probabilistic Incentivization

Though the discussion so far implies that current norms favoring the use of real stakes for experimental choices enhance the generalizability of experimental results, there is a counterpoint: these norms can have exclusionary effects on scholars who lack sufficient research funding (Bardsley et al. 2009). This limitation contributes to the overrepresentation of researchers 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).

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

One meaningful change in methodological norms that would decrease costs while still potentially preserving the external validity afforded by real stakes is disciplinary permission to use probabilistic incentivization. This involves (honestly) informing all participants that only a randomly-selected subset of their experimental choices will map onto real-world consequences, and/or that only the choices of a randomly-selected subset of the sample will map onto real-world consequences. 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 stakes conditions, and only report evidence of CHB (see March et al. 2016; 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 stakes condition 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 probabilistic incentivization or complete omission of real stakes.

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 incentivization. For participants in real stakes 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.

7 Conclusion

This paper shows that the recent hypothetical bias literature does not justify leaving experimental choices in most modern experiments without real stakes. 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.

Experimental economics' norms in favor of providing real stakes 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 stakes. However, the experimental economics literature is rich with examples where real stakes 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 stakes conditions may meaningfully impact experimental TEs, and thus to constrain experimental economists' choices by functionally requiring real stakes to be provided for experimental choices before experimental TEs are trusted. These real stakes be provided for all participants and all experimental choices, or for a randomly chosen subset of participants and/or tasks. What is important is that experimental participants make choices with the expectation that these choices may have real-world consequences, ensuring that behaviors observe in experiment are more aligned with the behaviors one can expect people to take in the real 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. (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 (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.

Andrews, Isaiah and Maximilian Kasy (2019). "Identification of and correction for publication bias". *American Economic Review* 109.8, pp. 2766–2794. DOI: 10.1257/aer.20180310.

Angrist, Joshua D and Jörn-Steffen Pischke (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.

Askarov, Zohid et al. (2023). “The significance of data-sharing policy”. *Journal of the European Economic Association* 21.3, pp. 1191–1226. DOI: 10.1093/jeea/jvac053.

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.

Brodeur, Abel, Nikolai Cook, and Carina Neisser (2024). “*p*-hacking, data type and data-sharing policy”. *The Economic Journal* 134.659, pp. 985–1018. DOI: 10.1093/ej/uead104.

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 No. 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. (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.

Chopra, Felix et al. (2024). "The null result penalty". *The Economic Journal* 134.657, pp. 193–219. DOI: 10.1093/ej/uead060.

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.

Fanelli, Daniele (2012). "Negative results are disappearing from most disciplines and countries". *Scientometrics* 90.3, pp. 891–904. DOI: 10.1007/s11192-011-0494-7.

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 (2024). *The need for equivalence testing in economics.* Institute for Replication Discussion Paper Series No. 125. URL: https://www.econstor.eu/handle/10419/296190.

Flood, Merrill M. (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.

Franco, Annie, Neil Malhotra, and Gabor Simonovits (2014). “Publication bias in the social sciences: Unlocking the file drawer”. *Science* 345.6203, pp. 1502–1505. DOI: 10.1126/science.1255484.

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

Harrison, Glenn W and E Elisabet Rutström (2008). “Experimental evidence on the existence of hypothetical bias in value elicitation methods”. *Handbook of experimental economics results*. Ed. by Charles R Plott and Vernon L Smith. Elsevier.

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.

March, Christoph et al. (2016). *Pay few subjects but pay them well: Cost-effectiveness of random incentive systems*. CESifo Working Paper Series No. 5988. DOI: 10.2139/ssrn.2821053.

Matousek, Jindrich, Tomas Havranek, and Zuzana Irsova (2021). "Individual discount rates: A meta-analysis of experimental evidence". *Experimental Economics* 25.1, pp. 318–358. DOI: 10.1007/s10683-021-09716-9.

Mosteller, Frederick and Philip Nogee (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.

Ortmann, Andreas (2016). "Episodes from the early history of experimentation in economics". *The making of experimental economics: Witness seminar on the emergence of a field*. Ed. by Andrej Svorenčík and Harro Maas. Springer, pp. 195–217.

Roth, Alvin E (1995). "Introduction to experimental economics". *Handbook of experimental economics*. Ed. by John H Kagel and Alvin E Roth. Princeton University Press, pp. 3–109.

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

Rubin, Donald B (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies". *Journal of Educational Psychology* 66.5, pp. 688–701. DOI: 10.1037/h0037350.

— (2005). "Causal inference using potential outcomes". *Journal of the American Statistical Association* 100.469, pp. 322–331. DOI: 10.1198/016214504000001880.

Schram, Arthur (2005). "Artificiality: The tension between internal and external validity in economic experiments". *Journal of Economic Methodology* 12.2, pp. 225–237. DOI: 10.1080/13501780500086081.

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 (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.

— (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. (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.

Wallis, W A and M Friedman (1942). “The empirical derivation of indifference functions”. *Studies in mathematical economics and econometrics in memory of Henry Schultz*. Ed. by O Lange, F McIntyre, and T Yntema. University of Chicago Press.

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.