NBER WORKING PAPER SERIES

EXPERIMENTS: WHY, HOW, AND A USERS GUIDE FOR PRODUCERS AS WELL AS CONSUMERS

Muriel Niederle

Working Paper 33630
http://www.nber.org/papers/w33630

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2025

I thank Emanuel Vespa for extensive discussions while thinking about and writing this chapter, and Shakil Rahman Ayan, Lukas Bolte, Nina Buchmann, Maxim Bakhtin, Arun Chandrashekar, Christine Exley, Tony Fan, Asia-Kim Francavilla, Kirby Nielsen, Maria Recalde, Alex Rees-Jones, Peter Robertson, Alvin Roth, Mart n Serramo, Sahana Subramanyam and Jason Somerville for suggestions, comments and discussions. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2025 by Muriel Niederle. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Experiments: Why, How, and A Users Guide for Producers as well as Consumers
Muriel Niederle
NBER Working Paper No. 33630
March 2025
JEL No. C9

ABSTRACT

This chapter is intended as an introduction to laboratory experiments, when to use, how to evaluate them, why they matter and what are the pitfalls when designing them. I hope that users as well as consumers will find Sections that broaden their views. I start with when an economist might want to run an experiment. I then discuss basic lessons when designing experiments. I introduce a language to start a systematic description of tools we have when designing experiments to show the importance or role of a new model or force in explaining behavior. The penultimate chapter provides an advanced toolkit for running experiments. I end this chapter with my views on pre-registration, pre-analysis plans and the need for replications, robustness tests and extensions.

Muriel Niederle
Department of Economics
579 Jane Stanford Way
Stanford University
Stanford, CA 94305-6072
and NBER
niederle@stanford.edu

Experiments:

Why, How, and A Users Guide for Producers as well as Consumers

Muriel Niederle[1]

Stanford University, NBER and SIEPR fellow

March 22, 2025

> "Nothing is less real than realism. Details are confusing. It is only by selection, by elimination, by emphasis, that we get at the real meaning of things." Georgia O'Keeffe[2]

This chapter is intended as an introduction to laboratory experiments, when to use, how to evaluate them, why they matter and what are the pitfalls when designing them. I hope that users as well as consumers will find Sections that broaden their views. I start with when an economist might want to run an experiment. I then discuss basic lessons when designing experiments. I introduce a language to start a systematic description of tools we have when designing experiments to show the importance or role of a new model or force in explaining behavior. The penultimate chapter provides an advanced toolkit for running experiments. I end this chapter with my views on pre-registration, pre-analysis plans and the need for replications, robustness tests and extensions.

In many economics departments, Experimental Economics is still(!) not a class that is readily available to graduate students. In this chapter I hope to provide an introductory remedy to this unfortunate state of affairs by starting with a basic understanding of when an economist might want to run an experiment, and if so, how they might go about it. Perhaps even more importantly, I also address how an economist might want to read experimental economics papers. This spans from an assessment about what aspects of an experiment are important, what worries one might have that turn out to be less of an issue in clean experiments, and which aspects one might want to keep being wary about; in short, how best to consume papers using this perhaps still somewhat exotic economic method. I hope that this chapter will help any economist from advanced undergraduate to even well-established colleagues in assessing the method of experiments. Since beginners in running experiments may have somewhat different "needs" than economists who only wish to – or find themselves in need to -- consume experiments, I will divide this chapter into appropriate sections which hopefully makes it clear which parts are especially relevant for whom.

[1] I thank Emanuel Vespa for extensive discussions while thinking about and writing this chapter, and Shakil Rahman Ayan, Lukas Bolte, Nina Buchmann, Maxim Bakhtin, Arun Chandrashekar, Christine Exley, Tony Fan, Asia-Kim Francavilla, Kirby Nielsen, Maria Recalde, Alex Rees-Jones, Peter Robertson, Alvin Roth, Martín Serramo, Sahana Subramanyam and Jason Somerville […] for suggestions, comments and discussions.

[2] I thank Lukas Bolte for this beautiful citation. It is taken from "I Can't Sing, So I Paint! Says Ultra Realistic Artist; Art is Not Photography—It Is Expression of Inner Life!: Miss O'Keeffe Explains Subjective Aspect of Her Work," New York Sun, December 5, 1922, quoted in Jonathan Stuhlman, Georgia O'Keeffe: Circling Around Abstraction (Manchester, VT: Hudson Hills Press, 2007), p. 22.

While I hope to convey general lessons, I will make them more accessible and understandable by providing specific examples. These will often, though not always, come from my own papers, or from economists whose work I know exceptionally well (mostly my advisors, students or coauthors). While this may seem self-serving, the main reason is that for those experiments I know– rather than have to infer– why authors made certain choices. And one aspect of experiments that will become obvious almost immediately, is that they require the researcher to make *a lot* of decisions. This chapter therefore is in no way a literature survey, nor really a highlight of amazing papers. It rather showcases papers whose history I am exceptionally familiar with. I will also not provide negative examples, but rather present potential pitfalls, with one exception in Section 3.1.

I start this chapter with some general views on when one might want to run an experiment. I discuss how experiments can be used, what options one might have (surveys, experiments or field experiments) and what the advantages and disadvantages of each method are. From now on when I say "experiments" I mean laboratory-style experiments, which is what this chapter focuses on.

In Section 2 I discuss basic lessons when designing experiments, such as the importance of theory or a conceptual framework (2.1) as well as potential pitfalls of being guided too much by theory (2.7). I discuss aspects to consider when designing an experiment (2.2 and 2.3) and the one rule you should really follow (2.5). In 2.4 I talk about an important final check after you finished designing but before you run your experiment. In 2.6 I discuss my view on pilots. I end the Section in 2.8 with a warning, a potential issue of any empirical work, *g-hacking*, which I think has not received sufficient attention. Basically, *g*-hacking (for game -- or situation -- hacking) is the issue that a researcher may try multiple games or environments before settling on the one that delivers the desired result, *while not disclosing this search process*!

In the third Section I introduce a language to start a systematic description of tools we have when designing experiments to be able to showcase the importance or role of a new model or force in explaining behavior. I discuss different strategies to control for alternative hypotheses, as well as their advantages and disadvantages. The main gist of the Section is that *background noise* or other models and forces may drive results rather than the model we are hoping to provide evidence for. *Background noise* is inevitable, as participants in the experiment play a specific game or act in a specific environment rather than in an abstract description of a game or environment. As such, for example, outcomes are in general paid in money rather than in utils. This on its own (together with many aspects present in the laboratory but not captured by the abstract general description) may trigger forces such as altruism (or salience, or complexity concerns) which may affect behavior.

The most important aspect of experiments is to ensure that the model you are focusing on rather than *background noise* or other models are responsible for the aspect of your results you want to focus on. In terms of accounting for specific alternate hypotheses (perhaps including the *background noise*) experiments have the advantage over just econometrics in that we can *design* rather than "just" estimate the "but for" or "alternative universe" when we eliminate various forces. This allows us to more directly infer their effect on behavior. I discuss the *design by elimination*, *indirect* and *direct control*, as well as old and new comparative static (do-it-both-ways) experiments. Finally, I am a proponent of *stress-testing* the hypothesis a final time before declaring victory. This is the longest Section, as I also include ample examples to showcase each methodology. I also provide evidence that many such design features, while perhaps easy ex post, are not always obvious ex ante, as whole parts of the literature sometimes took a while to adopt them.

Section 4 consists of a few more advanced and special tools that are handy to have in your toolbox though may not be discussed in all Experimental Economics courses. The last Section, Section 5, provides my take on pre-registration, pre-analysis plans, and the importance of replications and robustness checks.

I want to provide a final point of caution. While I believe that most experimental economists agree with my views, they are, in fact, just that, my views, albeit after lots of discussions and help from many students, colleagues and friends (it often does take a village, even just to write a paper). For everyone designing or consuming experiments: A great experiment is almost like a performance. It is supposed to look simple, obvious and effortless. This does not mean that it was easy, simple or quick to get there. I hope this chapter provides you with some guidance along the way.

1. When and when not to use an Experiment

I think instead of asking when to use an experiment, I'd almost turn the question around: Which questions aren't suitable for an experiment, really? In every science, laboratory evidence plays its part, and so, I think, it is, or rather should be, with economics. Though, of course, there are some questions for which experiments may be less suitable.

For example, when I write a theory paper, the method of choice is of course not an experiment. However, when I think about whether my model or paper has a chance to help us understand the world, I am often thinking about which data it could explain that other papers cannot, or what data I would need to see to determine whether the behavior of individuals is more likely to follow my rather than someone else's predictions. While sometimes I can think of field data, I am very aware that clean tests, in the end, are most often achieved by highly controlled lab experiments that more closely follow the assumptions of the model.

When I want to know whether a particular gender affirmative action quota in France can increase, over decades, the pool of women eligible for the quota, as well as the number of high performing women, experiments may not at first come to mind. There are two reasons for this. First, changes in the pool of high performing women may be due to a complex interaction of many mechanisms that may require years to take effect (as we show in De Sousa and Niederle, 2024). Experiments are rather better equipped at testing precise mechanisms. Second, experiments on a participant pool that differs from the one the researcher is interested in may not always be most convincing, especially when eliciting quantitative measurements. For example, Austrians enjoy watching skiing races, and, I am going to make a wild prediction, most will really prefer it to watching an American Football game. To assess the strength of these preferences, laboratory choices by American college students, or even on a representative sample of Americans, is probably not the most helpful evidence of Austrians' preferences. Note that this is a shortcoming of any measurement using a pool of participants that plausibly differs in their preferences from the pool of question (here, Austrians), whether it be in the lab or in the field.

So, I think Experiments are not helpful in showing the existence of complex interactions that may take decades, nor on estimating preferences using a participant pool that is very different from the pool of individuals one wants to study. However, even in many such settings experiments are still useful, we just must be careful how we interpret results and what conclusions we draw.

In my recent project with Lea Nagel and Emanuel Vespa (Nagel, Niederle, Vespa, 2025), we study why individuals fall prey to the winners' curse in first price common value auctions, a term coined by petroleum engineers (Capen, Clapp, and Campbell, 1971) considering returns from bidding in auctions for the rights to drill for oil. Kagel and Levin (1986, 2002) show that this is a robust phenomenon and even observed among bidders who have experience be it in the laboratory or outside of the laboratory. In our paper, we try to understand which forces make it difficult for individuals to avoid losses. While we think our insights are valuable beyond the laboratory, it would be foolish to think that our experiment has something to say how my Nobel-prize winning colleague Paul Milgrom would bid or what advice he would give to a company engaged in such an auction. However, an experiment might be helpful in understanding why *most* individuals (who are not Paul Milgrom) are not placing sophisticated bids.

Likewise, an experiment may be helpful in understanding potential mechanisms and drivers for *why* the gender affirmative action quota in France was helpful. I might turn not only to theory, but also to experiments, to investigate whether potential behavioral mechanisms are able to generate expected changes, since after all, experiments are a great way to gather data where they are lacking.

I think as economists, especially behavioral economists, we may underestimate the role experiments have played in advancing the field, be they experiments by economists or psychologists. I am even happy to entertain the, perhaps bold, claim, that a sizable majority of findings in behavioral economics have had their first appearance in the laboratory (though this may be the topic for another paper, one that perhaps needs writing to remind us of the role of experiments). As a literature, there is a growing push to provide evidence that phenomena in the laboratory are relevant to explain behavior outside of the laboratory. And, as a literature, this may have to be undertaken also by experimenters rather than waiting for other Economists to provide such evidence.

Overall, I hope this guide will also help clarify when an experiment would be particularly helpful, beyond helping you in preparing to run such an experiment. In 1.1 I discuss field experiments and in 1.2 surveys. In 1.3 I broadly discuss the concerns one might have with laboratory experiments.

1.1 Field Experiments: Pros and Cons

In contrast to laboratory experiments, field experiments live in a specific, rich, and sometimes complex environment. This generates specific scientific advantages as well as disadvantages. Field experiments also have a "perceived" advantage among many economists which I will address and question in this section as well. To conclude, I point out that the combination of both lab and field evidence is perhaps often the most conclusive. The advantage of living in a specific, rich and complex environment is that field experiments can address specific policy-relevant questions. For example, to study the returns of an extra year of schooling in the Dominican Republic, or the reasons for the wide-spread custom of child marriages in Bangladesh, field evidence and field experiments are going to be very informative. Clearly such questions are related to a specific rich environment.

Field evidence will also be more convincing when the researcher is interested in the preferences of a specific participant pool. For example, consider alpine Austrian farmers who, in the Spring, herd cows outside and to meadows often near the top of mountains (yes, it's not just the Swiss farmers who do that). Suppose I want to understand whether male and female Austrian farmers differ in their decision on which incentive scheme they would prefer to be used for their "performance" in "putting cows on top of the mountain," a piece rate or a tournament. Among Austrian alpine farmers, would women be less likely to select the tournament payment scheme than men? Experiments with American undergraduate students can provide some insights into gender differences in preferences for competitions (Niederle and Vesterlund, 2007). However, adult alpine Austrian farmers may behave very differently from American undergraduate students. If one truly wants to understand preferences and behavior of alpine Austrian farmers when herding cows to the top of the mountain, I would probably recommend running a field experiment with them.[3]

Likewise, if I want to understand the role of a "male breadwinner" norm in Bangladesh, and whether such a norm might impede the earnings potential of spouses, I may not want to use experiments of American male and female undergraduates and have them "pretend" to be a married couple in Bangladesh. So, when aiming to estimate level effects for a specific population of interest, such as the competitiveness of male and female Austrian farmers for a specific task, or aiming to understand potential earnings implications of norms on couples in a given society, experiments on undergraduates may not be the way to go. Even lab-style experiments that use the participant-pool of interest may not be the perfect method, if one worries that there are task-specific preferences, or special norms in, say, Bangladesh, that play a role when deciding upon careers rather than jobs of "just" a few hours, weeks, of even several months.

Field experiments have another advantage, that, however, I feel is more of an unfortunate state of affairs rather than a statement I subscribe to myself. They are often viewed by other economists as "more sexy," "more real," or the "gold standard." This is a statement I really do not subscribe to. While field experiments have a lot of plusses, especially in being able to gauge the exact level effect of a specific intervention in a specific environment, or the preference parameter of a specific population, they also have restrictions.

One important factor, especially when exploring new traits or behavior, or insights that are not geared towards a specific population or environment, is that laboratory experiments are in general much cheaper. Put differently, insights per dollar may arguably be much larger for laboratory than for field experiments. As such, especially when I venture into unchartered territory, this makes them for me the first choice. Another obvious restriction of field experiments is exactly what makes them also so appealing: They live in a specific and rich context. Let's go back to the Austrian alpine farmers with their cows who in the summer (the cows, that is) reside in alpine meadows. Around 2003, Lise Vesterlund and I were about to conduct the very first study on whether there are gender differences in preferences for a piece rate versus a competitive payment scheme. Should we have used alpine Austrian farmers and the task of "putting cows on top of the mountains"?[4] Even if a reader has heard of the task, I am going to make a wild guess that they do not know much about alpine Austrian farmers.

Compare such a field experiment to a laboratory experiment that uses a simple task – adding five two-digit numbers for five minutes – and American undergraduates. Lise Vesterlund and I provided simple instructions that explained the payment schemes and how to select among them. As a consumer and a producer, I am much, *much* more convinced by the findings of such a laboratory experiment where I can read and understand the instructions than of a field experiment on alpine Austrian farmers (even though I, at least, would be able to read those instructions as well). This is because we, as economists, through many papers, (somewhat) "understand" undergraduate students and participants in online platforms, know how they react in many environments, can gauge when results are perhaps unlikely.[5] Furthermore, compared to

[3] Though experiments with other participant pools may be useful in documenting how robust and ubiquitous such a gender gap is, see also Klinowski and Niederle (2024).

[4] Alternatively, we could have used a task that is common only among a small subgroup of people in Denmark, where Lise Vesterlund is from, which probably helps you figure out where I am from;).

[5] That being said, not everyone, to put it mildly, expected us to find gender differences, which was even more the case for my first paper on gender differences in competitiveness with Uri Gneezy and Aldo Rustichini (Gneezy, Niederle, Rustichini, 2003).

alpine Austrian farmers, we have a good idea that male and female undergraduate students who show up to laboratory experiments are otherwise quite similar.

In contrast, a consumer of the field experiment may have no idea about alpine Austrian farmers: Is that a domain reserved for men? Or are men involved in the more lucrative endeavor of running a restaurant/bar/hotel rather than occupying themselves with the few cows which enjoy a very healthy lifestyle ("happy cows") and likely are not a big part of the family business? The asset of field experiments, a specific and rich context, makes them also harder to interpret, especially when one is not completely familiar with the context or how it may interact with the main result (for a more detailed discussion of the role of context, see Section 3). That is, a rich context not only affects how plausible it is that the results are not driven by factors hidden in the environment, but also affects the external relevance of findings. Let me provide you with two more examples that show that once presented with a specific field experiment, we are sometimes aware of not only their advantages but also their limitations.

The first example comes via my Nobel-prize winning advisor, colleague and friend Alvin Roth. He, sometimes with me, studied unraveling in entry level labor markets, the idea that individuals receive job offers earlier and earlier from year to year in a chaotic market that is plagued with exploding offers, that is offers that require applicants to make decisions before having a chance to learn whether jobs they prefer have definitely ruled them out. The market of medical students seeking residency positions eventually, after years of unraveling, started using a centralized clearinghouse (for a summary of the history of the medical match see Roth, 2003, and Roth and Peranson, 1999). It turns out that judges who hire law clerks suffer from a similar problem. When trying to convince judges that for them, like for medical residencies, a centralized clearinghouse might solve the problem, Alvin Roth told me that the response was a polite rejection. (Think of something like "Professor, you do not understand, we are *nothing* like doctors!") However, an experiment using US undergraduates (who later may become doctors, lawyers or perhaps even economists) that shows that a centralized clearinghouse that uses a stable matching mechanism can halt unraveling was much more convincing to judges (though in the end, other forces prevailed). I think it is not unreasonable for judges to worry that a special feature of the medical profession and their organizations may have a strong influence on the operation of the market beyond just the matching algorithm of the centralized clearinghouse.[6]

In earlier work, aiming to establish the importance of stable algorithms, Alvin Roth (Roth, 1991) studied medical matches in the UK, and showed that the regions that kept using a centralized clearinghouse were largely those that used a stable algorithm. But this association was far from perfect. And, indeed, a doctor from the UK, looking at the evidence mentioned that the data are clear: Only doctors from Scotland and Wales like to participate in centralized clearinghouses, others prefer the more "natural" decentralized way of making and reacting to offers themselves. Contrast this to an experiment with two markets where the *only* difference between the two is the algorithm; one using a matching algorithm that is stable and the other one that produces matches that are not stable. John Kagel and Alvin Roth (Kagel and Roth, 2000) show that in this clean environment without added complexity, and specifically without one market using Scottish applicants and hospitals and the other using English ones, the market with a stable clearinghouse prevents unraveling while the other does not. This is a great clean data point that adds to the empirical evidence that was quite noisy and together helps confirm that what matters is likely the algorithm, since the algorithm on

[6] And indeed, markets with a stable centralized clearinghouse sometimes fail and start to unravel, see Niederle and Roth (2004) and McKinney, Niederle and Roth (2005).

its own *can* generate different results. As far as I know, John Kagel and Alvin Roth never tested whether Scotts are more inclined to participate in centralized matching mechanisms than the English are.

A second example where Economists understand that field data, potentially collected via field experiments, may "suffer" from overlooked "contamination," that is, generate (or magnify) results for reasons different from the ones directly studied, is development economics. Indeed, if we were to believe that individuals, norms, institutions […] in low and high-income countries were identical, there would be no need for development economics in the first place. Second, I think (and indeed hope) that a researcher who found that a specific intervention, or institutional change that has a big effect in country A (say Afghanistan) would not immediately conclude that the same effect would be evident in another country Z (say Zimbabwe).

While field experiments may suffer from the richness of the environment which arguably sometimes restricts how easily their results can be applied in other settings, they, in contrast to lab experiments, don't have many economists worrying about that, or that the results have *any* predictive power outside of the lab. This is often voiced as a concern whether lab findings have *external validity*. However, *external validity*, strictly speaking, is fulfilled if the result is found in a different laboratory with perhaps slightly changed instructions and a different environment. I think what Economists actually mean is *external relevance*, or really, *economic relevance*. The question is not (or at least should not be) *just* whether the finding can be replicated somewhere else, which speaks to their robustness. Rather, when Economists (mis-)use external validity, what they should really ask is whether the finding can account or speak to the respective economic question, that is the *external* and *economic relevance* of the results. Consider my paper with Lise Vesterlund on gender differences in tournament entry (Niederle and Vesterlund, 2007). Finding a similar result with alpine Austrians farmers still does not address whether competitiveness is an important trait that correlates with education and labor market outcomes nor whether gender differences in competitiveness can help account for the gender gap in those outcomes.

When the lab environment is very abstract and tests a specific mechanism or a specific preference or trait (such as competitiveness), only field evidence can determine the external relevance and most importantly the economic significance of the findings from the lab. In Buser, Niederle and Oosterbeek (2014) we combine a measure of competitiveness a la Niederle and Vesterlund (2007) to predict education choices of Dutch 9th graders which exhibit a significant gender gap. We find that competitiveness, above and beyond other information we have about the youths (such as grades, confidence, risk measures) predicts education choices and further, the gender gap in competitiveness accounts for about 20 percent of the observed gender gap in those choices. Likewise, the Kagel and Roth (2000) experiment showing the role of a stable algorithm when predicting whether a centralized clearinghouse is used rather than abandoned is more impressive in the combination with the earlier field evidence in Roth (1991).

Maybe, taking a step back, it is clear that the laboratory and data from a lab in the field (as in Buser, Niederle and Ossterbeek, 2014) or field data without an experiment (as in Roth 1991) all have virtues. I think as it is always in Economics, the correct answer to "What is the best method" is "It depends." And, perhaps most importantly, it is often when we combine different methods that they are, *each*, more convincing. As such, a goal of any good education should be that you can use the most appropriate method, rather being confined to only one way of addressing the research question at hand.

On a final note, when you have a question you want to study and wonder whether a field or a lab experiment is more appropriate, then, as a first pass, I would consider the following rule. If you are interested in levels, such as measuring a specific parameter for a specific population, then I would go to the field. If you want to test which mechanism is responsible for generating a specific outcome, then I would likely go to the laboratory.

1.2 Lab Experiments or Surveys: Pros and Cons

There are two major differences between laboratory experiments and surveys that apply in most (though not all cases): The amount of time they take for participants, as most – though by no means all surveys, just think of the long version of the census – are short. The second difference is that surveys do not contain payments that depend on the respondents' answers (apart from perhaps paying for giving an answer to all questions).

Surveys, therefore, are, by design, often cheaper. However, surveys that pay participants minimum wage to compensate for their time may not be much cheaper than experiments, though the latter in many states have to ensure minimum wage for each participant, which means average pay is often somewhat higher. Incentivizing choices is probably the biggest distinguishing feature of economic experiments. While almost universally used, there has been, and still is, a debate about the importance and role of incentives, and to what extent they are really needed. Personally, I am aware of two reasons when and why incentives matter.

The first circumstance when incentives matter is when other forces are present. For example, an individual who derives utility from holding certain beliefs will presumably be more inclined to distort their beliefs when incentives for accuracy are low. Alternatively, a participant may care about image concerns. The role of such other forces may increase without the competition of incentives. A nice example is in Carpenter (2005) who compares play in the dictator and the ultimatum game with and without incentives.[7] With incentives, the first mover suggests much more equal splits in the ultimatum game than in the dictator game. However, when the games are played without incentives, the proposals between the two games are much more similar. Presumably participants in unincentivized dictator games want to appear altruistic and, given the lack of incentives, can do so for free. In contrast, in the ultimatum game, the (justified) fear of rejection leads to offers close to the equal split even in the presence of incentives.

The second instance when incentives probably matter is when the task is one that requires substantial effort. When the participant pool consists of undergraduate students at prestigious universities and the task has some intellectual challenge, I would however not be surprised if participants perform highly even without incentives.[8] For participants who are less strongly motivated, either due to the participant pool not being students at elite universities who plausibly enjoy taking interesting tests, or due to the task itself being less inspiring, incentives may be important. Segal (2012) shows that a high performance on an unincentivized coding speed test (which is part of the US Armed Service Vocational Aptitude Battery, ASVAB) predicts future economic success because it measures intrinsic motivation rather than some kind of intelligence. It is her combination of field data and experiments using both incentivized and non-incentivized versions of the test that provides convincing evidence for this hypothesis.

[7] In both games the proposer (the first mover) splits a fixed amount, say $5, between themselves and another participant, the responder. In the dictator game the recipient has to accept the division of money, while in the ultimatum game the responder has the additional option to reject the division which results in a payoff of $0 for everyone (and the $5 reverting back to the experimenter).

[8] See Gneezy and Rustichini (2000) on how very small incentives may be detrimental, especially to performances of motivated participants, compared to no incentives or more substantial incentives.

There are only a few papers directly testing whether incentivized lab measures and survey questions are similarly informative. In Buser, Niederle and Oosterbeek (2024), we test the role of competitiveness on education and labor market in a representative sample of Dutch adults. We show *in the same data set* that a survey question on competitiveness correlates with an incentivized tournament entry decision a la Niederle and Vesterlund (2007) administered a year later.[9] More importantly, the two measures basically have very similar predictions on education and labor market outcomes, though they clearly do not measure the exact same thing. One is an abstract general question, and the other is a tournament entry decision for a specific task.

However, keep in mind that as experimenters our designs control for – to us – plausible alternative hypotheses, as well as alternative hypotheses many other economists may have. Given that many economists already worry that economic models may somehow not fully apply to experiments, I think it is, in general, prudent to use incentives, and be it just to avoid arguments that participants are indifferent and that we hence cannot trust their decisions.[10] For a (now slightly dated) meta-analysis and summary on when incentives matter, see Camerer and Hogarth (1999). Overall, I hope to see more efforts in understanding (and indeed showing) when incentives matter and when they do not, and why this is the case.

1.3 Cons of Experiments, final thoughts (and a final advantage)

Are there any drawbacks of laboratory experiments?[11] We already mentioned they may not be ideal when studying phenomena that rely on complex interactions that require decades to develop. Suppose, therefore, that there is a phenomenon, game or environment we can "squeeze" into a simple, somewhat short set of decisions and interactions.[12] Further, suppose that the agents we want to study are expected to select actions that maximize a given objective function. For a laboratory experiment to be useful in understanding why agents who are confronted with this objective function select specific actions, we need participants in the laboratory to, in essence, behave similarly to those outside of the laboratory. I use the nomenclature we developed in Nagel, Niederle and Vespa (2025) to describe three distinct classes of reasons why participants in the laboratory may deviate from behavior we expect to observe outside of the laboratory. Since we classify reasons why individuals deviate from a given objective function, this nomenclature is also useful in when addressing why individuals may deviate from the neoclassical prediction, which was our original motivation for this nomenclature.[13] I end this section with two comments: One is a potential answer when confronted with economists who are not engaging into which of those three possible issues are at play at a

[9] Specifically, the survey question is "How competitive do you consider yourself to be? Please choose a value on the scale below, where the value 0 means 'not competitive at all' and the value 10 means 'very competitive'."

[10] For a critique on flat incentives, see Harrison (1989), and for a discussion of this critique and how it does or does not apply in the specific context of auctions, see Section 3.2.

[11] Perhaps one con, or at least I think of it as a con, is that it is still a method that many economists are not that familiar with. This means that sometimes some of your very impressive design features may not receive the attention they deserve, and, likewise, some of your perhaps more questionable design choices will not be questioned as much as they should be.

[12] This requirement may look deceptively benign. Recall, for instance, that most real effort tasks in the lab are evaluated over a short time frame and, perhaps for this reason, have been found to be quite inelastic with respect to incentives. I think most economists would agree that for many tasks incentives do matter, though their effect may not be easily observable when it comes to short spurts of effort without much room for investment into proficiency at a task.

[13] Dellavigna (2009) introduces a somewhat different nomenclature with non-standard preferences, incorrect beliefs and systematic biases. Handel and Schwartzstein (2018) classify the failure to pay attention to information as frictions and mental gaps.

specific experiment, but rather voice that somehow what happens in experiments is "not real." The second is an advantage of lab experiments over any kind of field evidence that is often overlooked but I think scientifically very relevant.

1.3.1 Different Preferences

The utility or objective function of participants in the laboratory may be different from individuals outside of the laboratory, reducing the external relevance of laboratory findings. There are three reasons for such differences.

One reason might be that there are forces at play in the lab that are absent (or less pronounced) outside of the lab. Individuals in the laboratory know they participate in an experiment, which may exacerbate image concerns, including concerns about how they are viewed by the experimenter or experimenter demand effects.[14] Note that any such concerns may be exacerbated in a survey study, where there are no incentives that may dampen the costly behavior participants engage in just because they worry about their image. Furthermore, in the field, such concerns may be present as well whenever individuals are aware that they participate in an experiment.

Second, there might be forces at play outside of the laboratory that are absent in the lab, which might be responsible for generating different behavioral patterns in the lab than in the field. For example, coming back to image concerns, it could be that outside of the laboratory individuals have a reputation, or are able to document certain traits, both of which participants may not easily transport to the laboratory. An individual who cannot document their reputation may well differ in their choices from one whose reputation is well established.

Though we describe the presence of image concerns and the absence of participants' reputations in the laboratory as "bugs," they may also be features. For instance, one might want to study how strangers who have no reputation interact. This may be difficult if we are not able to pull participants away from the environment in which they operate in on a regular basis. While the laboratory may suffer from those "bugs," we can circumvent them, and as such also explicitly study their impact on decisions.

For example, concerning the observability of actions, there was a worry in the literature about reasons for the largely fair offers in ultimatum games. Do participants forego money by rejecting unfair offers to punish unfair behavior, and is this driving the prevalence of fair offers? Or, rather, are participants rejecting unfair offers and making fair offers because they feel that this is expected of them by the experimenter, and hence they engage in such behavior especially when in an experiment where their actions are not anonymous? To disentangle motives of participants, Bolton and Zwick (1995) introduce an anonymity treatment in which the experimenter is unable to attribute individual actions to individual participants and go through great length to ensure participants understand this. They find that anonymity somewhat increases the number of unfair offers, but only to a small extent. In contrast, when the responder of the ultimatum offer is prohibited from punishing unfair offers by removing the power to reject them, and hence turning the ultimatum into a dictator game, there is a substantial increase in unfair offers. Hence, the structure of the game, much more than the gaze of an experimenter, determines the prevalence of fair offers.

[14] That being said, anonymity outside of the lab may be less and less guaranteed as especially our online behavior may be much less anonymous than decisions we may do in laboratory experiments.

Similarly, the laboratory also allows us to study whether *increasing* observability of past behavior or "traits" of participants to perhaps more closely mirror opportunities outside of the lab, affects the participants' decisions.

Exley (2017) presents laboratory participants with a future volunteer opportunity that is observable to other participants, call them "judges." Consider participants who volunteer because they worry about their image, that is who want to signal their virtue. We might expect such participants to reduce the amount of volunteering for this particular charity if they have other means to document their reputation as individuals who engage in volunteering. To test this, Christine varies whether "judges" know about the participants' prior volunteer behavior. We expect participants who have a good reputation of volunteering, that is, have previously volunteered in the laboratory or outside of the laboratory based on self-reported behavior, to be affected by this manipulation. Christine finds that if the reputation of such a participant becomes known to judges, then the participant engages in less future volunteering.

A third and final reason why participants outside of the laboratory behave differently from those in the laboratory is that they have different preferences or utility functions. If the goal is to understand a specific group of individuals whose preferences we think are unusual (such as maybe gamblers vis-à-vis risk, nurses vis-à-vis altruism, Austrians vis-à-vis watching skiing races, etc), then using any other subsample may not be that informative.[15]

1.3.2 Different Constraints or Computational Problems

Even if participants in the laboratory are expected to have the same utility function as participants outside of the laboratory, it could be that laboratory participants have greater trouble at computing utility maximizing actions. Put differently, actions of laboratory participants may be subject to more significant frictions that result from computational costs and difficulties. Such frictions may distort findings in the laboratory vis-à-vis results from the field.

One reason for such increased frictions may be, as we already addressed, reduced stakes. Laboratory experiments can address this by varying stakes, for example, by moving experiments to low-income countries where stakes of several month's salaries can be implemented next to smaller stakes more commonly used in lab experiments.

A second reason for increased difficulties in the laboratory may be that individuals in the field have some expertise that results in them being better able at computing implications of various actions and hence finding the payoff maximizing one. I want to reiterate that, of course, it would be quite foolish – or heroic, depending on your predisposition – to assume that participants in the laboratory will bid in auctions like Paul Milgrom. However, I think it is much less heroic to believe that most individuals who participate in common value auctions, or in any transaction with an object that has some component of affiliated values, like buying a piece of art, say an Australian aboriginal painting or a house, do not behave like Paul Milgrom either!

There is a literature in experimental economics that considers whether "professionals" behave differently from "non-professionals" and standard laboratory participants. Frechette (2016) summarizes this literature with "Thus, it seems that in some cases, such as the sportscard traders with the endowment effect, behavioral

[15] Indeed, Frechette (2016) surveying the literature using standard and non-standard participant pools summarizes work showing that nurses, compared to average students, appear to be more altruistic.

anomalies are mitigated by experience.[16] However, taking the evidence of professionals as a whole, in many instances, what one would conclude from using students is qualitatively similar to what is observed with professionals."

1.3.3 Different Mental Models

A final reason in which choices of laboratory participants may differ from those in the field is when laboratory participants do not construe the problem or game or environment the way participants in the field do.

One reason is due to expertise, which we just covered, and which, as far as the evidence so far suggests, mostly does not generate a large difference in actions.[17] However, experts in the field may themselves behave differently than they do in somewhat abstract laboratory experiments. This could be for two reasons.

The first reason is that the field provided participants with a lot of feedback, and as such allows them to form heuristics that guide their behavior in a way to result in different behavior than in an abstract laboratory experiment. Note that, in principle, the lab allows for even more precise and controlled feedback, that is, we can, in principle, study the role of learning (which is indeed a large literature in Experimental Economics.) Furthermore, I want to emphasize, that any such "strict" usage of a heuristic that does not adapt to changes in the environment – like a setting in the laboratory – will result in potentially misconstruing behavior as following the neoclassical model. Specifically, such participants cannot – as their failure to adapt to the lab shows – be expected to behave similarly optimally when the environment changes! And a lot of economics is interested not only in describing behavior, but also predicting choices as we change the environment.

A second reason, very related to the first, is that not only "extreme experts," but even many "standard" laboratory participants are sometimes stymied by abstract representations. The most famous example of this is perhaps the Wason selection task (Wason, 1966). Participants in an experiment are told there is pack of cards which each have a letter on one side and a number on the other side. Four cards are taken at random from the pack, two showing letters, of which one a "D", and two showing numbers, of which one a "3" and one, say, a "7". The experimenter tells the participant to turn over only the necessary card to verify whether the cards fulfill the following rule: If there is a "D" on one side of a card, then there is a "3" on the other side. Most participants turn over the "D" card, but some also the "3" card, and very few the "7" card.

In contrast, when the task is described using perhaps more familiar concepts such as destinations and mode of transport (Wason & Shapiro, 1971), and famously underage drinking: If a person is drinking beer they have to be at least 21 years old (Griggs & Cox, 1982), then participants are much better at checking the person who drinks a beer and the underage person, while leaving the older person "alone."

I think it is perhaps the Wason selection task which sometimes leads experimenters to use "naturalistic" environments. Personally, I in general advise against this unless there is a clear reason why it is important,

[16] However, adding to the short conclusion of Frechette (2016), it is noteworthy that sportscard enthusiasts who buy cards at trade shows and are in the "natural habitat" do exhibit behavior that mirror results from the laboratory.

[17] However, and I cannot emphasize this enough, I would not expect laboratory participants to behave in common value actions like my friend and colleague Paul Milgrom. However, I am at the same time very happy to entertain the notion that most bidders, even some who engage auctions involving large sums of money, are not Paul Milgrom. Furthermore, if you engage in a history project, I hazard that before Paul Milgrom's Nobel-prize worthy work, no one, perhaps including Paul Milgrom, was bidding like Paul Milgrom would now.

as it is in the Wason task. The reason is that complications that turn out to be needless and just for the aesthetic complicate the design. In Section 3 I discuss at lengths the considerable advantages of clean and clinical designs.

1.3.4 Final Comments on Lab Experiments

Why do many Economists prefer field evidence to laboratory experiments? I believe many Economists feel that field data or field experiments have more external validity. It is as if they feel that Economic models apply everywhere, when we make high stakes decisions like career choices, buying a house, marrying, and when we make low stakes decisions, like buying milk, detergent or breakfast cereal (a vibrant literature, it turns out), but worry that Economic models do not apply in this little separate universe called the experimental laboratory. Put differently, it is as if the validity of Economic models is akin to the extent to which the Romans (validity of Economic Models) conquered Gaul (the world) in Asterix the Gaul (Goscinny and Uderzo, 1961). The English translation (Goscinny, Uderzo and Bell, 1969) reads:

> "The year is 50 BC. Gaul is entirely occupied by the Romans. Well, not entirely... One small village of indomitable Gauls still holds out against the invaders. And life is not easy for the Roman legionaries who garrison the fortified camps of Totorum, Aquarium, Laudanum and Compendium..."[18]

Put differently, it is as if some economists worry that Economics is valid in all our lives, apart from this one little corner called the experimental Lab…I think those Economists underestimate the power of their models. I do believe that they apply to undergraduates making decisions in well described and incentivized environments, even if this environment happens to be the experimental laboratory. I do not believe that our economic models are so fragile that we experimenters can make them "not work" whenever we want. If we do not believe in the power of economic models in the lab, I think we should worry what else economic models (or, as in my analogy, the Romans) failed to conquer? Simultaneously, I think Economists attach too much mystery to economic laboratory experiments, though I love the idea of us Experimenters having a magic potion that transforms each of us into powerful heroes (as the Gauls in the small village in my analogy).

There is a final, often very overlooked factor, that, ceteris paribus, is in favor of laboratory experiments: replicability. I think Experimental Economics, as a subfield of Economics, has been especially good at promoting replications. Those replications are rarely "exact replications," or papers written just to replicate a result, rather they happen naturally in papers that study extensions and variations of the original finding. A researcher who studies the effect of a variation reruns the main treatment of the original study and compares the results of *their* replication of the main treatment using their experimental procedures, language, participant pool […] to the outcome from their extension or variation. Hence, beyond the study of the variation, as an aside, the paper generated a replication of the main treatment, using often different parameters, procedures, participants… For more on the value of replications, see Coffman and Niederle (2015) and Coffman, Niederle and Wilson (2017).

[18] "Nous sommes en 50 avant Jésus-Christ. Toute la Gaule est occupée par les romains... Toute? Non! Un village peuplé d'irréductibles gaulois résiste encore et toujours à l'envahisseur. Et la vie n'est pas facile pour les garnisons de légionnaires romains des camps retranchés de Babaorum, Aquarium, Laudanum et Petibonum..."

2. Basics when Designing an Experiment

In an earlier chapter (Niederle, 2015), I wrote about the answer to the question “What is a good experiment?” “My first reaction is to answer my empirically minded colleagues ‘Well, let me ask you: What is a good regression?’ Clearly a good experiment (or regression) is one that allows testing for the main effect while controlling for other plausible alternatives. This helps ensure that the original hypothesis is reached for the right reasons and the initial theory is not wrongly confirmed. However, there is an aspect of experimental design that is probably closer connected to theory than empirical work: As designers, we are responsible for the environment in which the data are generated. As such a good experimental design also needs to fulfill requirements one may impose on good theory papers: The environment should be such that it is easy to see what drives the result, and as simple as possible to make the point.”

I am still happy with this answer, though in this chapter, I will expand on how to approach designing an experiment. When designing an experiment, there are basically no fixed rules (apart from one, see 2.5. This is both a feature and a bug, as it requires the researcher to make many decisions, see 2.2, and, in addition, to be forward looking in thinking how to analyze the data (see 2.4 and what control treatments to run, see Section 3. I establish the importance of a model or conceptual framework (2.1) but also that relying too much on a model may backfire (2.7). I discuss my views on Pilots (2.6) and end the Section with an overlooked issue, g-hacking in 2.8.

I start with the most basic way to approach designing an experiment, and I leave some more advanced techniques to Section 3.

2.1 Stuff Happens vs Competing Hypotheses

The most important and first item when designing an experiment is to know what the goal of the experiment is. Experiments, apart from measurement exercises, which I do not cover here, are best described as using data to provide evidence for a specific hypothesis. In general, what this really means is to provide evidence *against* a host of *alternative* hypothesis.

Sometimes researchers start with an interesting environment that is either taken from observations, theory, or general curiosity, and then design an experiment to “see what happens, what participants will do.” In the beginning of my career, when experimental economics was still less established, some researchers would generate such environments, sometimes without even knowing the predictions of a rational agent. Such “one treatment experiments” will yield results, since participants must do something. Which is why I call them “stuff happens” experiments. However, the outcomes are then often hard to interpret, especially if the researcher would be able to provide a possible explanation for many different possible outcomes.

In general, when designing an experiment, it is very important to have a hypothesis about the expected behavior, as well as an alternative hypothesis, and then design the experiment such that the “dream outcome” allows to clearly reject the alternative hypothesis (See Section 2.4)

2.2 A Long Way to a Million Choices

Designing an experiment is often a long process of back-and-forth between what one wants to test and hopefully be able to convincingly show, as well as thinking about a simple and clean experiment. I personally, in general, start with a hypothesis, but then I often have a phase where I think more in the “design” rather than the “model” space. It is as if the design gets a life of its own, where lots of simplifications, variations and controls come to mind. Then, I move back to the model or hypothesis space to see what I really want to be able to say at the end, what explanations I have to consider. I check what each piece of the design contributes, or whether some pieces are just complications that are not necessary. As with good writing, where a main advice of Strunk and White (2000) is to omit needless words, so it is with good design: Only keep what you really need, as everything else can potentially affect results in ways you are not aware of. And then, the whole process restarts, quite a few times, actually.

The most important aspect of designing experiments is not to get to enamored of a particular design piece too early, and always be able to answer, "Why did you make this choice?" The answer shouldn't be: "Good question, I don't know." Ideally it is "Well, if I had chosen these two other ways, here is the problem of each way." However, sometimes the answer is "Well, we had to make a decision between these options, all seemed good, but we had to pick one." This is because you will have to make *a lot* of decisions, many of which are not controversial, but since you design the whole environment, you still have to make them.

For example, in Martinez-Marquina, Niederle and Vespa (2019) we show that individuals have a hard time to think of hypothetical compared to realized events. The final environment of our experiment consists of a buyer who can purchase a firm that is either of low (say 20) or of high (say 120) value, both equally likely. The buyer does not know the true value of the firm v, submits a purchasing price p, and has profits of $1.5v-p$ if $p \geq v$ and profits of 0 otherwise.[19] We find, as is common in this "acquiring a company" environment, that individuals are likely to submit a price of $p=120$ and hence select a 50:50 lottery of -90 (when the firm has value 20) and 60 (when the firm has value 120) rather than a price of $p=20$ which results in a 50:50 lottery with payoffs of 10 and 0, respectively.[20] If the problem consisted of only one firm of known value v, we would expect (hope) many (all?) participants to submit a price of $p=v$ and receive maximum profits of $0.5v$.

The main idea of our paper is to point out that there are two distinct changes or complications that are removed when moving from the case of a firm with two potential values to the case of a firm with a certain fixed value. One is to move from two values (or states or contingencies) to one value, the other is to move from hypothetical (or uncertain or potential) values to realized (or deterministic) values.

To decompose underlying causes of mistakes into these two components we introduce a treatment that requires the participant to think of two values (or contingencies) but without the added uncertainty. In the deterministic treatment, there are two firms that both exist, one of low and one of high value. The participant submits *one* price that is then sent to each of the two firms. Hence, the participant can, dependent on the price, buy no firm (if $p<20$), one firm (if $20 \leq p<120$) or two firms (if $120 \leq p$) and hence have profits of 0, 10 or -30, for $p<20$, $p=20$ and $p=120$, respectively. If for each dollar in the deterministic treatment the participant receives two dollars in probabilistic treatment, then, for each price p, the certain payoffs in the deterministic treatment are equal to the expected payoffs in the probabilistic treatment.

Put differently, we wanted to provide evidence that a significant portion of the mistakes in the acquiring a company problem are due to participants having to think through hypothetical contingencies (like in the standard probabilistic treatment), which are of course absent when there is only one firm of known value.

[19] Think of a story that the buyer is a better manager of the firm and hence, if they buy the firm, can increase profits by 50%, and the current owner sells if they receive their value of the firm.

[20] In addition, those same participants would all select the 50:50 lottery of 10 and 0 rather than the one of -90 and 60, confirming that submitting a price of 120 is not a case of a sudden mass-outbreak of desire to be risk-seeking.

However, it could well be that conceptual and computational complexities of having to consider two states by themselves are the root cause for mistakes. This is why we created the deterministic treatment where participants also have to think about two states or two firm values, the low and the high firm, engage in appropriate computations, but where values are realized rather than hypothetical.

Alejandro Martinez-Marquina, Emanuel Vespa and I were designing an experiment to show that hypothetical states cause problems to participants that are not present when states are realized, and hence problems that are not due just to purely computational complexity. Around January or February of 2017, when Emanuel Vespa was visiting Stanford for a year, we had a finished design that involved a game and individuals, basically, not foreseeing equilibrium effects. We wanted to show that individuals have a hard time thinking about the actions of the other player, and that this may be easier if other players had chosen their action already. It was important to, however, not reduce the complexity of the game otherwise. Basically, in our main treatment (the probabilistic treatment) the participant was playing against someone who, say, had a 50% chance to move left and a 50% chance to move right. The deterministic version that turns hypothetical events into deterministic events had the participant play simultaneously in two games, specifically they could only select one action that then applied to both games. In one game, the other player had moved left, and in the other right. And then everything was a bit more complicated to consider equilibrium effects as well. I do not fully remember the setup, because, when discussing the project with colleagues and other visitors, Georg Weizsaecker asked us why we needed a game rather than an individual decision-making environment. We realized that he was right, that the only reason for a game was that we ourselves fell under the curse of path-dependency: We started thinking about failures to predict equilibrium effects, refined it to failures to think about hypothetical states, but kept the environment of a two-player game.

However, a game is only needed when thinking about the lack of participants to foresee equilibrium effects, but not when showing that realized states are treated differently from hypothetical states. To have a clean design, do not be afraid to revise, revise and revise your design again. Consider the advice that your paper should not be a history lesson, that is, should not show your trains of thoughts and your discovery process of the main result. In general, it doesn't matter how you thought about the issue at hand, your readers want to see the best way to get your main result, rather than a way that is personally meaningful to you as it shows how you discovered and solved the problem. And just like your paper, neither should your experimental design be a history lesson of your thought process, be ruthless in eliminating now superfluous design aspects.

What makes designing experiments feel like an exercise in making a million choices is that after you have thought of all the pieces of your experiment, you still must design and choose how to implement them. You have to select specific numbers, think about incentives, write code and instructions. All these tasks take time and I can only recommend that you recruit friends in trying out your experiment and providing you with feedback, specifically if you're not a native speaker (like me). A good, clean, crisp design is a bit like a magic trick or a performance: It is supposed to look effortless and simple, but do not be deceived. You are not seeing the beginning but the end of an often long process!

2.3 Kinds of Experiments most Suitable for Beginners

There are broadly speaking two kinds of experiments. The first, call it an "any news is good news" experiment, is one where the starting point is that there are two competing models that make opposite predictions in a specific environment. The goal of the experiment is to provide a data point to the question as to which model makes (more) accurate predictions in this specific environment.

For example, when I was working with Ned Augenblick and Charles Sprenger, there were two "camps" on time discounting. One camp I kind of belonged to (by introspection and well, after all, David Laibson was one of my advisors) posited that present bias, or hyperbolic discounting is something many individuals suffer from.[21] The other camp can be summarized as pointing out that the (at the time) vast majority of existing experimental evidence of present bias from "money now versus money later" is flawed. The flaw criticized by Andreoni and Sprenger (2012) is that some of those papers have individuals decide between money now (or somewhat better, a check now) and a check dated for a later time t' compared to a check for some later time t and a check at t+t'. Even individuals without present bias, but with some costs to cashing checks (or in better versions, chance of forgetting to open the mail to receive future checks or losing checks they cannot cash immediately) would prefer money or a check now even though this may come with a cost. Money now versus checks later experiments would then attribute this cost which reflects the chance of losing mail or future-dated checks as the reduced value of certain money in the future compared to money now. Hence, they would find evidence consistent with present bias even for participants who do not suffer from present bias.

Andreoni and Sprenger (2012) in a clean design controlling for the effort of collecting the money, albeit with an inbuilt delay even for immediate payments, fail to find substantial evidence of present bias.[22]

In Augenblick, Niederle and Sprenger (2015) we point out that the model of present bias is one of consumption, and that for the standard experimental participant pool, gaining a little bit of money may not really alter their consumption – though a few people may be happy just to "receive some money." In our experiment, Ned Augenblick, Charles Sprenger and I compare choices in the money domain to ones in the consumption domain, in our case effort. Ex ante, I was guessing that we probably replicate the failure of evidence of present bias in the money domain (even though in our paper immediate money and later money is always delivered as cash in hand). In addition, I thought we would find evidence of present bias in the consumption domain. Outcomes of finding present bias in both domains would have still been interesting, while results of failure of present bias in both domains would have made me, at least, a bit more hesitant about present bias.

As such, the experiment in our paper was an experiment where there was almost a clear way of how to run it, though we had some innovative design aspects. Additionally, it was a case where essentially any findings or news were good news (though, as is always the case, some good news are better than others). Such an

[21] Basically, a present-biased participant discounts payoffs received tomorrow compared to payoffs received today by $\beta\delta$, but discounts payoffs between some future time t and t+1 by only δ. That means, when deciding how much to pay to receive payoffs just one day earlier, the participant would pay much more if the earlier day is today, rather than any in the future.

[22] Specifically, participants receive a check with some amount both at time t and t+t' in their mailbox. They receive an additional sum of money they can divide between those two checks, where, for this additional money, the experimenter imposes an exchange rate between money at time t and at time t'. A participant who does not discount at all is going to put all their money where it is worth more, that is either all at t or all at t' (unless the exchange rate is 1:1 between these two points in time). The main variation is whether t is today (in which case the decision is between an immediate and a later payment) or one week or a month from now (in both cases all payments are in the future). They find that divisions of money were basically unaffected by whether t is now or later. A proponent of present bias may worry that a check received in the late afternoon is not exactly "money now" and as such may question whether present-bias motives even had a chance to be relevant.

experiment is especially good for beginners, as they are in some obvious way safer, and also, conditionally on finding such a scenario, easier. This is because the amount of innovation needed to design such an experiment is often lower than in the style of experiments I discuss next.

The second type of experiments, which I call "finding new truth's" experiments, are experiments where most economists expect a specific outcome, while your intuition is that something else will happen, something not predicted by any established model. As it happens, the vast majority of my experimental papers fall in that domain, though of course not the paper Ned Augenblick and Charles Sprenger. Such papers are much scarier for two reasons.

First, they often require a completely new design and environment. This is because if you were to use standard designs, we probably would already know about your new insight. Designing experiments far from existing experiments is difficult, and something that may require some expertise. This is not a sign of the result to be volatile and is in fact the case even if your insight and hypothesis is very robust. For example, consider my first paper on gender with Uri Gneezy and Aldo Rustichini in 2003. We showed that the gender gap in performance significantly increases when moving from a piece rate to a tournament incentive scheme. The task we used was to solve as many mazes as possible in fifteen minutes. This was one of the first real effort tasks which are now both more common and often involve much shorter time frames of five minutes or less. As it turns out, many real effort tasks in the laboratory are not very elastic, that is, performance does not react much to incentives. If we had used one of those alternate tasks, then, even though the hypothesis is true, we would not have been able to confirm it. This is because for most tasks there is no change in performance for either men or women when moving from a piece rate to a tournament payment scheme. Hence, there is no chance to detect a gender difference in the change in performances across payment schemes. It turns out solving difficult mazes is a task that responds to incentives, for both women and men. Therefore, this is an environment that allows us to test whether women, whose performance does respond to incentives, do, however, not increase their performance as much as men do when moving from a piece rate to a mixed tournament incentive scheme. Basically, a task where performance does not react to incentives generates an environment where we cannot meaningfully test our hypothesis. In general, finding an environment where your unusual hypothesis can be tested is not easy when you need the environment to enable specific traits and preferences to manifest themselves. However, Uri Gneezy, Aldo Rustichini and I did manage to find such an environment, and this was the beginning of my work on competitiveness.

The second reason such "finding new truth's" experiments are scary is that, even if you're a good enough experimenter to construct an environment where you can study your hypothesis, if your intuition is wrong, and you do not get your result, the paper is basically super uninteresting. My advisor, Alvin Roth, who was, and still is, extremely supportive, didn't expect me to find gender differences in competitiveness, and I bet so did most economists, so, not finding such results, would not have been interesting.[23]

For this reason, I really do not recommend any new experimenter to start their career with a "finding new truth's" experiment. Even if such an experimenter has a great idea and intuition, it is hard to generate the environment that allows to confirm the special insight. This may be especially the case if the researcher relies on specific psychological traits or preferences to manifest themselves. For example, while we all agree that there are numerous pleasant activities for individuals outside of the laboratory, it is surprisingly

[23] On that note, this also suggests that not all null results are equally interesting, or informative.

difficult to generate a "fun" short task, where we can measure the performance, and in Augenblick, Niederle and Sprenger (2015), we spectacularly failed.[24] Such "finding new truth's" papers are a high variance endeavor, and graduate school and tenure clocks are on a fixed schedule. So, maybe it is wise to not just aim for such papers. That being said, many of my papers fall into that category, and some have led to fantastic outcomes.

For all these reasons, and this may look self-serving, I recommend a young experimenter to start working with a more senior experimenter early in their career. You will receive more credit for coauthored work if you don't always coauthor with the same senior person, and once you have your own well-published papers. I, myself, learned a lot from more senior coauthors. And I invite you to ask my students or anyone who has worked with more senior coauthors in case you worry this is too self-serving an advice.

2.4 Dream Outcome

While designing experiments is already a long (really long) process of back-and-forth thinking, there is a final step that is important to do. Once settled on a final design, I advise my students to think of the following: What do the dream data look like, and can they answer the questions you have, the hypotheses you want to test? If the data turn out to be different, which other treatment would you run? While it is always good to be optimistic, thinking in advance about potential future control treatments that may have to be run is important. This is because some of the million choices that were made in designing the experiment, many of which may not be important for the main hypothesis, may be important when designing a control treatment. If you didn't think, in advance, what such a control treatment might look like, well, then some of the previous choices may turn out to be very unfortunate – and for no good reason, as this state of affairs could have been avoided with some advance thinking.

However, after all is said and done, there are two kinds of researchers: The first kind think too little in advance, and then run many treatments, because their earlier decisions were mistakes. This only becomes a little problematic if they do not report all those early treatments, an issue I return to in Section 2.8. However, it is also possible to think too long, as at some point, hypothetical thinking *is hard*, after all we run experiments because we do not necessarily know what will happen! The most extreme case, not of thinking too long, but not knowing what will happen was in my project with Ned Augenblick and Charles Sprenger, where we three had different beliefs about what would happen in our experiment (Augenblick, Niederle and Sprenger, 2015), what a fun collaboration that was. I was also surprised by results in my recent project with Lea Nagel and Emanuel Vespa, which I will describe in more detail in Section 2.6.

2.5 One Rule specific to Experimental Economics

There is one, and really only one rule when running experiments that I would not recommend an experimenter, and especially one starting out, to engage in: Deception. The basically universally accepted rule is that everything that is mentioned in the instructions has to be true.

Note that I have heard about a much stricter view on deception, like never using data for reasons other than described while they are elicited. I do not adhere to this stricter view. In fact, I have violated that rule often,

[24] We had two real effort tasks, one, transcribing blurry Greek letters, which we expected to be costly to participants, and it was. The second task was a modified Tetris game. We were hoping for this to be a fun task, so we could study time preferences for costly as well as pleasurable effort. However, our modification of Tetris seems to have taken all the fun out of it, and participants found the task almost as costly as transcribing blurry Greek letters…

as it is one of the great design features we have at hand for within-person experiments. For example, in one of my early papers with Lise Vesterlund (Niederle and Vesterlund, 2007), we use the piece rate performance from round 1 to influence payment in round 4, and participants were not aware of this fact in round 1. In my most recent project with Lea Nagel and Emanuel Vespa we elicit strategies in round 3, strategies which are then used in rounds 5 onwards, though participants were not aware of this in round 3.[25] From now on, deception is defined as deviating from what you told participants in the experiment.

Mostly, not allowing deception is just a constraint that makes experiments more expensive than they would be if deception were allowed. For example, in a dictator game, if a participant is told that they have to divide $20 between themselves and the recipient, this recipient has to be another person who may have to be paid just to collect money without necessarily themselves producing any data. As such, eliminating that recipient might be very tempting. However, a proposer who decides how much money to keep and how much to give to the recipient, may act very differently if it were known that the other participant is, in fact, the experimenter, who simply takes some of their experimental money back (which is an additional reason the researcher might be tempted to engage in such deception).

Why are experimenters, in contrast to, at least historically, psychologists, “obsessed” with deception? I can think of two reasons, both boiling down to the interpretation of data from participants that may not trust that the instructions which explained how their choices translate to payments are followed and hence true.

One worry is that a participant who does not trust instructions behaves more erratically, and as such “injects” noise. While for individual decision-making experiments this only increases costs as more data points are needed to obtain results, a substantial number of noisy participants may be more problematic in group experiments where participants interact with each other. For example, if in a market game with induced demand and supply (see Smith, 1976) some sellers are willing to sell at a price below costs, and buyers are willing to pay more than what the artificial object is worth to them, then this changes the competitive equilibrium of the market. If Economists are more likely to have experiments with interactions in perhaps large groups than Psychologists do (which I think was definitely the case before the advent and prominence of online experiments), then it makes sense that Economists care more about eliminating noisy participants.

A second worry, which applies even in individual decision-making experiments, is that participants who do not believe the instructions may not only be more noisy, but also behave in a *biased* way. For example, consider the classic Ellsberg paradox experiment where participants prefer 50:50 bets to bets of either red or black when red and black balls are drawn from an urn of an unknown distribution of red and black balls.[26] This is evidence of ambiguity aversion.

However, it could also be evidence for the participant not believing instructions and not trusting the experimenter and worry the experimenter manipulates the composition of the ambiguous urn (perhaps throwing in a bunch of yellow balls that always lead to a loss).

In the laboratory, sometimes such a distrust in what the experimenter says and does can be mitigated by, for example, having one participant, at random, be the “randomizer” who first decides what balls to put into

[25] For a summary of views on deception, see Charness, Samek and van de Ven (2022).

[26] There are two urns of 100 balls each, Urn R(andom) consists of 50 red and 50 black balls, while urn A(mbiguous) contains an unknown mix or red and black balls. Individuals are in general indifferent between betting on red or black in the R urn, where betting red (black) pays $X if a red (black) ball is drawn from the urn, and $0 otherwise. The classic result is that individuals prefer to bet on red *and* on black in the R over the A urn.

the urn, and later draws balls from the urn. Once experiments are run online, this is of course much more difficult, and as such it is perhaps even more important to maintain the rule that there is no deception in economic experiments.

Even in the laboratory, if participants do not believe instruction or are routinely deceived, they may worry that some players are "confederates" which may, for example, lead them to play an ultimatum game like a dictator game if they think confederates always accept any offer.

The reader may correctly argue that if there is *never* any deception, what would be the harm of sometimes using it? While I am in general open to all kinds of experiments, I would not recommend breaking this rule as a young experimenter, and even a senior experimenter better have a really good reason for doing so. Indeed, at a conference I participated in, a senior economist (not a known experimenter himself) used deception and was not able to finish his talk. The issue is that modern institutional review boards in general require experimenters to disclose any deception at the end of the experiment to all participants. Hence, an economist who uses deception spoils the pool of future experiments. As such, many economic laboratories (though of course less psychology labs and not online settings) have a strict rule of either forbidding deception experiments outright or requiring that experiments that use deception announce that possibility in large font to the participants (which, as someone once complained, may take the "fun" out of using deception.)

2.6 Pilots

The advent of online experiments increased a practice that already attracted attention 30 years ago: Running extensive pilots. I almost never run designated pilots.

What are the pros of using pilots? Sometimes an experiment relies on specific initial results or some specific environment. For example, if I run an experiment to decompose the winners' curse in common value auctions, there better be a winners' curse to begin with.[27] Similarly, in Gneezy, Niederle and Rustichini (2003) we studied whether there is a change in the gender gap in performance when moving from a piece rate to a tournament payment scheme. It makes sense to test whether there is *any* change in performance when moving from a piece rate to a tournament payment scheme, that is, whether the performance is elastic vis a vis incentives. Indeed, many tasks do not exhibit such elasticity, in which case it will be impossible to find such a change in the gender gap in performance, even if it were present in all tasks where performance is elastic!

Put differently, pilots are great in testing whether the environment may suffer from floor or ceiling effects: To observe gender differences in tournament entry in Niederle and Vesterlund (2007), it better be the case that the tournament is attractive enough that some participants enter the tournament, but not so attractive that all participants enter the tournament (summarize this perhaps as "variance is your friend").

However, what researchers sometimes envision when running a pilot beyond the previous reasons is a pilot that provides an initial test of the hypothesis. In principle this can be done in an experiment that is otherwise not yet ideal and knowingly not the final version of the experiment. The main advantage is that a researcher

[27] Given the robustness of the winners' curse (Kagel and Levin, 2002), finding an environment in which participants do not suffer from the winners' curse may be interesting per se, but a different kind of experiment, warranting a different set of treatments than one aiming to decompose the winners' curse. So far, though, this is not an issue that has ever happened.

may quickly learn whether their initial hypothesis has empirical validity without spending a very long time to design a clean experiment around this hypothesis.

There is, however, perhaps especially for less experienced researchers, a huge potential cost of running such pilots. There are three possible outcomes, either the pilot is "successful" and shows that the initial hypothesis has empirical validity, or it is "clearly unsuccessful" and shows there is "not much" there, or it has some weak findings but perhaps some surprising side results. I will tell you why all three outcomes can make life difficult for a young researcher.

When the pilot is neither "successful" nor "clearly unsuccessful" but has some "side results" that look interesting, this can lead to the temptation to take these side results too seriously. Then researchers, based on a shitty (in a technical sense) pilot abandon their initial idea and start "chasing" significant results. But if the experiment was not designed for those, it is not clear how seriously such results should be taken, especially based on a smallish pilot. This can lead to a chase down a rabbit hole without the chance of a rabbit at the end, however. Note that, to be sure, I do *not* want to imply that researchers should not "listen to the data" and see where they lead. Nor do I deny that we often learn things from experiments we may not have fully anticipated – that's why we run them in the first place, to know, and learn. I *am*, however, concerned when such insights are formed in a sometimes panicky way based on small pilot data, just because the researcher is desperate to believe that there is *something* interesting there.

Almost worse, sometimes, is the case when the pilot is "successful." While this was perhaps only intended as a sign that there could be a positive result, the pilot then sometimes generates a life of its own. This can lead to the researcher now only thinking of designs that are close to the pilot, since that "worked." Instead, a great design should be tailored to the question at hand and the relevant alternative hypotheses that should be controlled for. I have observed that, especially for young researchers, it is hard to abandon a design that is not perfect but that "worked in a small pilot." This can lead to final designs that are not as good as they could be if the researcher had taken the pilot as what it is, a good sign, rather than a quick design they are now shackled to.

Finally, suppose the "shitty" pilot is "clearly unsuccessful." Does the researcher conclude that this means the hypothesis is not correct? Since, after all, the design of the pilot was not thought through and perhaps just for this reason results in noisy data that do not confirm the hypothesis. This then often leads to a second, maybe only slightly less "shitty" pilot, and the process restarts.

For all these reasons, I am very nervous when my students want to run pilot studies, and I often discourage them, especially if they are run too early. I do, however, acknowledge that using pilots properly can be very helpful, as contingent thinking is hard, and it may help the researcher understand what they ought to control for in their experiment.[28] If a researcher understands the value of a pilot, ex ante and ex post, and will not be shackled by them, they can be informative and useful. However, the rigorous practice and (excessive) use of pilots can generate a whole new host of problems which have been mentioned in Roth (1991) and which I will address in Section 2.8.

[28] In field experiments, pilots are probably more valuable, as they are often a mixture of testing for floor and ceiling effects, that is testing that the researcher found the "correct environment" to even test their hypothesis in the first place. It becomes more problematic when there is a large component of "testing for the main hypothesis in a small sample." Even in field experiments I would advise to have a very, very advanced design ready before starting to do pilots.

All that being said, in my most current project with Lea Nagel and Emanuel Vespa we, in the end, report our first two experimental treatments in the appendix, as if they were enormous pilots. We had two treatments, and some initial data on a third treatment, that we thought would be the paper. When we were discussing whether to finish gathering the data for the third treatment, we had the glorious insight that, given everything we now know, we can now run an even better, fully within-subject design, that will allow for more additional analyses than the ones we did so far. We decided to do that, and spent a few weeks redesigning and programming our new "mega-treatment." We then held our breath as the data came in, in sessions of 21 participants at a time in almost two-hour laboratory session, until we had run all 10 sessions for a total of 210 participants. The results were as expected, almost exceeding our expectations, and all our previous results were replicated. A few months later, presenting the new data, a famous researcher and editor asked, quite in disbelief, why we did that, whether a referee asked for it. We had (and as I am writing this, have) not yet submitted the paper, we didn't even have a draft yet, we just felt that we can do quite a bit better, and since the paper has our names on it, we wanted it to be a better paper! This may not be the best strategy to write as many papers as possible, as we probably could have submitted the early version and then redo the paper in response to referees, but this isn't what we did, we tried to turn this paper into one of our "big" papers. Note that, in passing, we are therefore even more certain of the robustness of results.

2.7 Theory is a Guide but can also be a Hindrance

Theory, or well thought through hypotheses are almost essential when designing an experiment, even if the theory is more a conceptual framework, otherwise, one may fall prey to designing a "stuff happens" experiment. When designing an experiment to test a theory, it is important that the theory applies in the first place. So, if the theory asks for common knowledge, make sure the relevant statements are at least public knowledge.[29]

A good experiment, however, may also need controls which, in a strict sense, are not necessary under the model. In fact, being too much guided by theory may lead to both design mistakes and erroneous conclusions.

Perhaps the best example how theory could lead us astray comes from the project with Lea Nagel and Emanuel Vespa on decomposing the winners' curse where we show how a "classic" design relying on theory can lead to erroneous conclusions.[30] One starting point of our paper is to point out that after 40 years of research and a host of behavioral theories to account for the winners' curse, no one had yet tested for the perhaps somewhat pedestrian possible explanation that bidders may overestimate the expected value of the item conditional on having the highest signal. This simple mistake could, in theory, account perhaps even for the whole winners' curse. Specifically, while bidders may bid more than the Bayesian expected value of the item conditional on having the highest signal, they may overestimate this value and bid less than their inflated valuation. Alternatively (or in addition) bidders may be subjected to a host of other forces that push

[29] See Roth and Murningham (1982) for the role of private versus public knowledge in bargaining games, and Huang, Kessler and Niederle (2024) on the importance of information when predicting the role of fairness in bargaining games.

[30] In common value auctions each bidder receives a signal about the common value of the item and places a bid that is guided by the bidder's beliefs about the expected value of the item conditional on winning. A common result is that in first price common value auctions the winning bidder has the highest estimate of the item, as expected in a symmetric BNE, but fails to sufficiently shade their bid and hence ends up winning the auction at a price that leads to expected losses, hence the name winners' curse.

their bid away from the symmetric risk-neutral Bayes Nash equilibrium. These forces range from unusual preferences such as risk-aversion or, maybe more exotic, a joy of winning or anticipated regret, to general computational problems and finally unusual mental models that break the relationship between winning and having the highest signal, such as cursed equilibrium (Eyster and Rabin, 2005), level-k thinking (Crawford and Iriberri, 2007) or a failure of contingent thinking (Niederle and Vespa, 2023). To hone in on the problem of computing the expected value of the item conditional on having the highest signal, we design bidding rounds where, when submitting a bid, bidders also provide an estimate of the expected value of the item if their signal was the highest among all seven bidders.

We find that bidders significantly overestimate the expected value of the item conditional on having the highest signal. Furthermore, while the majority of bidders bid more than the true expected value of the item conditional on having the highest signal, they bid *less* than their inflated estimate to the expected value of the item. If we rely on the symmetric equilibrium in first price common value auctions, we conclude that for most bidders the main culprit of falling prey to the winners' curse is an inflated computation of the expected value of the item. However, this conclusion is naïve and wrong if bidders do not fully, or not at all, condition on having the highest signal when submitting a bid. While this is the case in the symmetric equilibrium, it may or may not be fulfilled. And indeed, unusual mental models such as cursed equilibrium (Eyster and Rabin, 2005), level-k thinking (Crawford and Iriberri, 2007) or a failure of contingent thinking (Niederle and Vespa, 2023) break the relationship between winning and having the highest signal!

In Nagel, Niederle and Vespa (2025), for the first time, we directly assess, via an experimental design, whether participants, when submitting a bid, condition on having the highest signal. This is in contrast to using bids to infer whether individuals likely did or did not condition on this event. We find evidence against the hypothesis that bidders largely condition on having the highest signal. This also puts into doubt the conclusion that the inflated estimate of the expected value is the main culprit for the winners 'curse. This is because this conclusion rested on the, as we now know, mis-specified model that participants, when placing a bid, condition on having received the highest signal (as they would in a symmetric Bayes Nash equilibrium). For details see Nagel, Niederle and Vespa (2025) and discussions in Sections 3.4 and 3.5.

To summarize, using equilibrium analysis would have led us astray in the conclusions we can draw from the fact that (i) bidders overestimate the expected value of the item conditional on having the highest signal, and (ii) bidders often bid *less* than their inflated estimate of the expected value of the item.

Other reasons why relying too much on theory when designing experiments can lead us astray are given in Section 3.2. This Section points out that substantial changes between two environments that, using neoclassical theory, only boil down to a single change, may lead to misattributing changes to the single neoclassical reason, while instead they are driven by some of the other changes which, using neoclassical theory, do not matter, but, in fact, may matter a lot.

I think most experimental Economists are aware of the dangers of relying on narrow views about which model affects behavior. I will even go so far to claim that Experimenters, perhaps more than any other applied economists, are very attuned to the fact that a myriad of forces are at play when individuals make decisions, be they in a laboratory experiment or outside of the lab. Perhaps this is our true superpower, or the result of the magic potion that Getafix, the druid from Asterix's home-village (or Panoramix in French, or Miraculix in German, the three languages I have read Asterix in) provides us.

2.8 G-hacking

Experiments that aim to provide evidence either in favor of an unusual model or against the neoclassical model use a (potentially very small) set of games or environments in which they hope to make their point. However, often they do not explain how they selected the game or environment. Did the researcher select the game or the parameters at random, or did they unconsciously make a selection that favors their hypothesis? More extremely, did the researcher consciously look for games or environments that favor their hypothesis, either by using their intuition, or by doing extensive pre-testing and piloting? From now on I will use the word game, as a placeholder for all situations or environments or individual decision making tasks, and indeed, those feature heavily in the examples in this Section.

With "*g-hacking*," for game-hacking, I want to denote the practice of selecting a specific game (or specific parameters) or a specific environment, perhaps using extensive piloting, *while at the same time* presenting results *as if* the game was "randomly" drawn from a larger set of possible parameters and games. A researcher who is g-hacking presents their potentially very special results while declaring that they pertain to a whole set of games (or parameters). Put differently, suppose a researcher finds a result in a specific game with a specific set of parameters. Consider two possible scenarios how the researcher selected the game.

- In scenario 1, the researcher tested many versions before settling on the one game they present in their paper.
- In scenario 2, the researcher selected a game structure and then randomly selected parameters to use.

Clearly, as a consumer of the experiment, we would, ex ante, be much more likely to believe that the findings from the experiment represent likely outcomes in this large class of games if the results are from scenario 2, the random game, than from scenario 1, the heavily g-hacked game.

I denote the practice of scenario 1 as g-hacking, as trying to hack one's way through multiple games before finding the one that works. G-hacking is and should be reminiscent of p-hacking (Simonsohn, Nelson and Simmons, 2014), where g-hacking is deciding *which* data to collect and from which environment, and p-hacking pertains to how many of those data to collect, and which of the many possible analyses to report.

I will first discuss in which instances carefully selecting a game is not problematic. The second example concerns a case where authors may have engaged in g-hacking, even if only at a subconscious level. The third example concerns a case where the careful selection of games and environments of early papers lead to a biased literature. I end this section with a proposal on how to avoid unconscious g-hacking – I hope every reader refrains from deliberate g-hacking which includes hiding that they carefully selected the environment after extensive piloting. I also add at the end a warning that g-hacking may not have received the attention it, I fear, deserves.

An example where authors carefully select the environment and are quite open about it is from Richard Thaler whom I have heard talk how he and Kahneman and Tversky were trying out a lot of scenarios before they homed in on some of the now famous scenarios of "the beer on the beach," or "Linda the bank teller," or "the jacket and the pocket calculator"… The examples are presented and viewed as what they are, namely carefully selected scenarios that provide evidence, a proof of concept, if you will, that there are situations where behavior is not likely to conform to the neoclassical model, so-called anomalies. These examples do not appear to be particularly contrived or special, but it is clear that they are just that, examples. They, however, rightfully convinced Economists that there is something extremely interesting going on. And indeed, the concepts they represented were picked up in follow up work by many different experimental and behavioral Economists, and many have entered main-stream economics.

Carefully selecting an environment or game and being open about this is, however, not how many papers are written. Often researchers just present results of a specific game without mentioning why they chose these particular parameters. An example where authors may have engaged perhaps only in subconscious g-hacking is from the early days on two player constant sum games. Some researchers argued that behavior of laboratory participants are close to Nash equilibrium, and this is indeed what they found in their experiments. Others argued that participants do not play Nash, and indeed, in their experiments, they did not.[31] Just from those few examples, *where we do not know how they were selected*, we do not know whether players almost always or perhaps only very rarely use strategies that correspond to the Nash equilibrium!

Erev, Roth and Slonim (2016), analyzing previous studies, found that, among the games with unknown selection criteria, the farther the equilibrium is from 50:50 choices over the actions, the less likely participants are to play equilibrium. There are, however, a few caveats to this result. First, it results from the behavior in a small number of games with different procedures used by different researchers (they basically acknowledge the potentially substantial role of *background noise*, see Section 3). That is, we do not know whether differences between observed and predicted behavior generalize beyond these particular games, or whether observed deviations in each experiment were rather due to the specific features of the experimental procedures.[32]

To make direct comparisons feasible, Erev, Roth and Slonim (2016) vary the games they study. They consider as a starting point the class of two player two action constant sum games that have a unique non-trivial mixed strategy equilibrium. How should they select games from this (very large) class of games? One way is to ensure a nice “spread” in some metric, like for example the difference between equilibrium play and equal choice of actions. This would have the advantage that it could provide the best chance to detect changes in behavior as that metric changes. However, they note that a disadvantage of this approach is that the researchers still select specific games or the metric over which to have a nice spread of games, which could, in itself, bias the selection of games. Therefore, to avoid any chance of an unconscious bias, they randomly sample from all possible games. As it happens, their 10 games still provide a wide spread in the difference between equilibrium play and playing each action with the same chance. In addition, they employ the method discussed in Section 4.2 to make payoffs linear in probability and hence can safely assert that participants who follow the neoclassical model should play according to the unique mixed-strategy Nash equilibrium.

With those changes to previous experiments, they find that the closer equilibrium play is to 50:50 play in actions, the closer the behavior of participants is to equilibrium play. That is, their findings are consistent with the previous evidence from different experimenters using different procedures and different

[31] Some authors, like O’Neill (1987) got a step further. On page 2106 he writes why his particular game should receive more weight than other evidence. “A problem in empirical research has been the design of an experiment that accurately tests the theory. Here I describe an experimental game chosen to avoid two previous difficulties. First, the game allows calculation of the solution without assumptions about the exact shape of the players’ utility functions for money. Second, the game is easy for the subjects to comprehend; in fact, it is unique in being the simplest nontrivial game possible according to a definition of simplicity to be given in Section 3. My subjects’ behavior was close to minimax, and I suggest that this confirmatory evidence should weigh strongly against past failings of the theory, on the grounds that the design used here is more appropriate.”

[32] On that note, many papers did also not consider risk preferences when using Nash equilibrium to predict behavior.

parameters. In fact, they find that this relationship is stronger than that found from comparisons across papers, suggesting that the different methods applied by different researchers introduced noise which weakened the strength of the relationship that determines whether participants play according to the Nash equilibrium prediction.

Erev, Roth and Slonim (2016) provide one method to avoid g-hacking. Another method to not oneself engage in even unconscious g-hacking, and convincingly provide evidence against conscious g-hacking is to follow the literature. However, when many researchers follow "canonical" or established environments, this can generate a new kind of danger. Sometimes, the literature "forgets" that initial discoveries were made with potentially very carefully selected parameters. Such specific examples then sometimes create "a life of their own," are often extensively replicated, showing their robustness. However, the literature then sometimes forgets that this robustness for a given set of parameters or games is not equivalent to a robustness of results across a large span of parameters. If many papers are written using similar parameters, or similar games, we may vastly overestimate their robustness to only seemingly innocuous variations.

Let me give you an example of a very recent paper that highlights this issue and questions the perceived wisdom. There are two phenomena discussed in the paper. The first is the common consequence effect, known as the Allais paradox (Allais, 1953): Suppose you present participants with the following pair of choice tasks, where the residual probabilities give you no payment.

- Problem 1:
    - Lottery A: 100 percent chance of $1M
    - Lottery B: 89 percent chance of $1M and 10 percent chance of $5M
- Problem 2:
    - Lottery C: 11 percent chance of $1M
    - Lottery D: 10 percent chance of $5M

Note that lotteries C and D are created from lotteries A and B by changing the common consequence of an 89 percent chance of $1M into an 89 percent chance of $0. Many participants however make different choices in the two problems, and opt for option A and option D.

Allais (1953) also introduced the common ratio problem, with the canonical example coming from Kahneman and Tversky (1979). As before, it involves two pairwise choices.

- Problem 1:
    - Lottery A: 100 percent chance of $3000
    - Lottery B: 80 percent chance of $4000
- Problem 2:
    - Lottery C: 25 percent chance of $3000
    - Lottery D: 20 percent chance of $4000

Lotteries C and D are created from lotteries A and B by scaling down the probabilities of the non-zero outcomes by a common ratio of 0.25. Once more, many participants opt for options A and D.

Note that expected utility theory predicts that both the common consequence and the common ratio variations should *not* impact the relative preference for the two options. Both of these examples have been replicated many times (though not with millions). These examples were also fundamental in influencing the development of new behavioral models on risk that deviate from expected utility. These models, then, make predictions on a large set of parameters (being useful models) which would lead us to expect the common consequence and the common ratio violations to be common for a large set of problems and parameters. But how common are they, and are the results really as robust as we would expect?

McGranaghan et al. (2024b), analyzing data from previous meta-studies (Blavatsky et al, 2022, 2023), show that over 30% of prior common ratio experiments used the canonical parameters from Kahneman and Tversky (1979) and over 40% of prior common consequence experiments used the canonical parameters from Allais (1953). Furthermore, many studies find substantial non-expected utility behavior at and near these parameters. In contrast, the effects are significantly smaller, and sometimes even reverse, for other parameters (see also Jain and Nielsen, 2024).

To have a clearer sense as to the role of parameters, and where we would expect to find a common consequence and a common ratio violation, McGranaghan et al. (2024) collect new data that systematically cover the parameter space. They find that preference patterns consistently vary with the parameters of the decision problem. Specifically, the canonical non-expected utility effects do not always generalize, with even the opposite patterns emerging in certain areas.

So far, I mentioned studies that openly are very selective about the game or environment they consider, or studies that speak to a set of existing work and show that a literature suffered from g-hacking, whether done consciously or unconsciously.[33] How, then, can a researcher who is aiming to show a new phenomenon avoid g-hacking? While one can obviously refrain from extensive testing and piloting to find an environment that "works," we might still have to be careful not to overuse our intuition or insights and conduct a "poor researcher's" version of g-hacking, by thinking about what outcomes we expect from multiple scenarios rather than using costly pre-testing. However, complete protection from a "poor researcher's" g-hacking may, to some extent be impossible. However, we can take steps to reduce the extent of potential "insight" g-hacking we engage in.

In my work, I try to be as open as possible why I picked certain scenarios and discuss what I expect to be robust. For example, in Nagel, Niederle and Vespa (2025) we picked a canonical environment, rather than invent our own version of a common value auction. We used a generic environment since we wanted to decompose the winners' curse, and not have a design that homes in on one model while neglecting others. Rather, we wanted to provide equal footing to various models, and as such use a very canonical environment, or at least do so as much as we were able to. Note that we, however, just previously showed that using canonical environments may come, as a literature, with its own problems.

Furthermore, sometimes the point of an experiment is to come up with a new and clever environment which took a long time to think of and which has specific properties in controlling for a bunch of potential hypotheses. This is of course valuable as well, and something I have done, like the quite unusual common value auction in Ivanov, Levin and Niederle (2010). In this latter paper, we highlighted the amazing feature of this unusual environment that allowed us to specifically test certain models.

[33] The reason a literature suffers from g-hacking even if first papers in the case of Allais etc do not, is that the literature at some point forgets its roots and claims that the results from the studies on those selected games or situations are robust to not only those selected parameters but to many parameters that extend beyond those often considered in the literature.

When I do not test specific theories, as we did in Ivanov, Levin and Niederle (2010), but rather show a new phenomenon which we hope is more general, we often use multiple games from a class of games. We then try to select parameters randomly among a specified set of parameters to not unconsciously select some that push results to go one way or the other. That is, even for papers that do not speak to a broad existing literature we can employ methods just like the one used in Erev, Roth and Slonim (2016).[34]

Sometimes, selecting random parameters for a given set of games is not feasible, like in some of my papers on gender norms. We then try to follow clear structures on how we select scenarios. For example, in Dean, Exley, Niederle and Sarsons (2025) we used all gendered questions taken from the General Social Survey and the World Value Survey that were asked regularly and recently, rather than cherry-pick questions or make up questions. In Dean, Exley, Klinowski, Niederle and Sarsons (2025) we basically use all measures from the economics literature that have a very well documented gender component. We are aware that relying on previous studies does not eliminate g-hacking (see the discussion surrounding McGranaghan et al., 2024). However, at least the reader does not have to worry that we engaged in *even further* g-hacking than has potentially happened in that whole literature already.

I think that as experimental and behavioral economists, and as consumers of such experiments, we have to be more careful in acknowledging the dangers of g-hacking and the value of studies that systematically vary parameters or games. If the goal is to show an example, a proof of concept, unconscious g-hacking is perhaps almost unavoidable, though presumably more fruitful than the drunk man who searches for his car keys under the light. Conscious g-hacking, while tempting to hide, should be reported! Furthermore, note that meta-studies are not always helpful in addressing the generality of findings. This is the case when g-hacking has been rampant and when there is a strong tendency for follow up studies to use parameters close to those used in previous studies, as documented by McGranaghan et al. (2024).

The potential and the danger of g-hacking is especially large when a paper wants to make a general claim about the validity of some models. This is especially the case when a researcher wants to perhaps argue that their model is a better predictor of the data than some other model. It then better be the case that this is not only true for the g-hacked games you selected! As my advisor and friend Al Roth told us: “Theories are not like toothbrushes. They should work and be useful on other people’s data as well, not just your own.”

I think that g-hacking has not received sufficient attention as a questionable research practice. John, Loewenstein and Prelec, (2012) address the problem of selective reporting on collected data (see also Roth, 1994).[35] G-hacking is basically the step before that, that is deciding which studies to run, rather than ex post deciding which to leave out of the paper. And while it does not sound as bad as leaving out whole experiments, it is still very bad in terms of guiding research more than we might be entitled to, especially if we want to claim “random” selection rather than extensive pre-testing.

With the rising prevalence of online experiments g-hacking is perhaps especially “easy” as extensive piloting is really quite cheap and does not significantly reduce the participant pool for future studies. However, g-hacking may also be a problem in the field. A researcher may extensively pilot ways to gather

[34] For example, I employ this method in Fragiadakis, Niederle and Knoepfle (2022) and Martinez-Marquina, Niederle and Vespa (2019). In Kessler, Kivimaki, Litwin, and Niederle (2024) we employ two canonical games with lots of parameter constellations akin to McGranaghan et al. (2024).

[35] John, Loewenstein and Prelec (2012) ask about the prevalence of: In a paper, selectively reporting studies that “worked.”

data or try out many different environments. If they then go on collecting data where it “worked,” this may be the case for legitimate reasons, but it could also be for spurious reasons that do not pertain to the hypotheses the researcher considers, which means the result is prone to being questionable as the environment was g-hacked.

3. How to Design Experiments and Test Hypotheses

In this chapter I describe several strategies on how to provide evidence for your main hypothesis. Basically, suppose you have data about a behavior in an environment or game that suggests your model rather than the neoclassical model is a better fit to the data. In Section 3.1 I discuss why such evidence may fall quite short in being convincing evidence that your new model is a better predictor of behavior than the neoclassical model. Basically, the problem is that many other forces may be at work, partly due to your design or implementation of the game or environment. I call all these other forces *background noise*. I discuss how to design a new treatment to test whether the *background noise* or your new model is instrumental in generating the pattern of behavior you observed. In a *design by elimination*, you find a new game or environment that eliminates the force of your new model while keeping the environment as similar as possible such that forces from the *background noise* are as similar as possible across the two games. If the behavior in the new game or environment is very different from the original one and importantly does not generate the pattern from before, then you showed that an alternate universe where your new model has no predictive power substantially changes behavior. That is, if we believe that all effects of the *background noise* are identical across your two games or environments, then you provided evidence that your new model or force is essential in generating the original behavior.

In Section 3.2 I discuss old school comparative static experiments that compare behavior across two games, but with less attention paid to keeping the environment among the two scenarios as similar as possible. I discuss when this is less of an issue, namely if the goal is to provide evidence against the neoclassical model (and as such even forces in the *background noise* may serve as legitimate alternative forces). This method is less popular these days, as in the modern era of experimental economics the goal has shifted from providing so-called anomalies towards identifying new forces and models.

In Section 3.3 I discuss the case where there are several models or forces that may affect behavior, and your goal is to, as much as possible, exclude the impact of those forces to be able to conclude that your new model is responsible for the results. The logic of a design by elimination can easily be adapted to multiple alternative models. However, sometimes control treatments that serve as the alternate universe eliminating your new force are hard to come by. To make progress in such a case, a popular method is to employ what I call *indirect controls*. The first step is to aim to measure a key parameter of the force to be excluded or controlled for (like, for example the risk preference of a participant, or their computational capacities). The second step is then, in general, to control for the role of this force econometrically. I provide a detailed example that makes clear what the (often hidden) assumptions are of such a method of indirect controls. These assumptions are of two kinds: The first set concerns the justification for a specific way to measure the force to be controlled for. For example, risk preferences or computational capacities can be measured in different ways, potentially each muddled by a different *background noise*, which is often ignored. The second concerns the less unusual assumption that the underlying model used in the econometric exercise is correctly specified.

In Section 3.4 I discuss an alternative way to control for several forces. Basically, exploit the advantage experiments have over just using econometrics: We can design rather than have to estimate the “but for” or alternative universe when we eliminate forces to more directly infer their effect on our results. To do this, we adapt the insight from a *design by elimination*. Instead of finding an environment that eliminates the new model or force you want to hone in on, you aim for an environment that keeps the new force present but eliminates the force you aim to control for: a design I call a *direct control*, as it relies on fewer assumptions. That is, a direct control of a “nuisance” force like computational problems or risk preferences is an environment as similar as the original one, but where this force has no predictive power. This can be achieved by turning the environment into one where participants are risk-neutral, or computations are done for them, all that while keeping the environment as similar as possible to not generate new background noise. I provide specific examples where indirect controls would have led us to the wrong conclusions, and how direct controls were able to point us towards the correct model of behavior.

In Section 3.5 I provide a new technique which is best described as a *new comparative static*, or a *do it both ways* design: Basically, sometimes it may be really difficult to generate an environment that is similar to the original environment, but where you can argue that a specific force does not affect behavior, be it a “nuisance” force or your new model. In this case, you can aim for a less “extreme” version of a design by elimination or direct control: Find a new environment where the relative impact of the force differs from all other possible explanations. For example, say you have an environment where participants largely select the A rather than the B button. Suppose you generate a new environment sufficiently similar to the old one to arguably not change the effect of the *background noise*, where all forces would predict behavior to stay at the A button, but your model predicts a switch to the B button. If participants change to the B button, then this is evidence in favor of your model. I provide an ancient and a more recent example of such a new comparative static experiment.

In the final Section, Section 3.6, I provide an argument for stress-testing that the model or force you want to push is indeed responsible for the outcomes. Such a stress-test consists of an auxiliary prediction that would be true if your conclusions are correct.

3.0 Some Language and Definitions

An abstract Game $\boldsymbol{G}$ is defined as containing all aspects relevant for the model, that is the set of players, their possible strategies and how their strategies determine outcomes which are in utils. Note that I will talk about games in this Section in a broader sense, that is I also include one-player decision problems. For simplicity, from now on, I will always use the word game, even for individual decision-making tasks, or any kind of environments or situations. The description of an abstract game may differ depending on your model. For example, a neoclassical complete information normal form Game $\boldsymbol{G}^N$ contains the set of players, their actions, and how their strategies determine outcomes which are in utils. The neoclassical model m^N is the equilibrium concept (or solution concept for cooperative games) that determines the behavior we expect from participants.

When you have a new force that you think is a better predictor of behavior, there are broadly three way this can happen, using the nomenclature we developed in Nagel, Niederle and Vespa (2025) that I described previously. Players may have different preferences, different constraints and computational problems, or different mental models. In the case of different preferences, this implies a change in the description of the abstract game. For example, the model m^* may be that players are altruistic, that is their payoffs depend not only on their outcome but also on that of other players. As such m^* may have its own description of a Game $\boldsymbol{G}^*$ that has the same number of players, the same actions, but differs in how actions determine outcomes. In contrast, in the case of constraints (or computational difficulties) or different mental models, the difference is not so much in the description of the abstract Game, but rather in the solution concept. For example, suppose you believe that players act according to the mental model described by *level-k* thinking, $m^{level\text{-}k}$ in contrast to the neoclassical Nash equilibrium m^N. The description of the abstract Game $\boldsymbol{G}$ is identical between those two models, that is $\boldsymbol{G}^N = \boldsymbol{G}^{level\text{-}k}$, but the predictions by the model m^N and $m^{level\text{-}k}$ differ. This is because under the neoclassical model beliefs about the behavior of opponents have to be correct in equilibrium, while this is not the case under the level-*k* model $m^{level\text{-}k}$. In the *level-k* model, a level-0 player is, in general, supposed to play randomly, a level-1 player best responds to a level-0 player, a level-2 player best responds to a level-1 player and so on.

To fix ideas, I provide a concrete example of different preferences. The $\boldsymbol{G}^N$ version of the ultimatum game (Güth, Schmittberger, and Schwarze, 1982) with players who have a linear utility function in money consists of a fixed amount M player 1 divides into two parts, *x* for themselves and *M-x* for player 2. Player 2 can accept the division, providing player 1 with x and player 2 with M-x, or reject the division, providing both players with nothing. That is the acceptance of the division leads to payoffs (x, M-x) for players 1 and 2, respectively. In contrast, if m^* is a model where players have a linear utility function in money but are altruistic in a specific manner, namely that their utility consists of the minimum of each players payoff, then the acceptance of the division leads to utils (min{x, M-x} and min{x, M-x}) for players 1 and 2, respectively.

Suppose you run an experiment on some specific problem or *game g* that could be the representation of a game in either $\boldsymbol{G}^N$ or $\boldsymbol{G}^*$. In the game *g* which you implement in the laboratory, outcomes are in general in money (and not in utils). That is, the game *g* is the situation participants experience in the laboratory. The actions and information provided during the game *g* is compatible with the notion that the game *g* is a representative of either Game $\boldsymbol{G}^N$ or $\boldsymbol{G}^*$ in terms of actions and information and how actions turn into monetary consequences (though not into utils). This is because in game *g* participants are paid in monetary terms and not in utils.

Furthermore, the game *g* will have to select a specific amount of money M to be divided, which may be a small amount of $1, a medium amount of $10 or a large amount of $1000. Furthermore, to implement the game *g* in the laboratory you have to make decisions on aspects the description of the abstract Game $\boldsymbol{G}^N$ or $\boldsymbol{G}^*$ is silent on. For example, is the game *g* played using the strategy method or do participants play “live,” do participants make offers anonymously via the computer or face-to-face, how much do they know about the other participant, and where does the money comes from, is it via a money drop from the experimenter, or maybe because one or both players via their actions contributed to the pie? Any of those choices made when implementing the ultimatum game in the laboratory may (and as we now know do) affect behavior. That is, there are many forces at work that are not included in the abstract description of the Game $\boldsymbol{G}^N$ or $\boldsymbol{G}^*$.

Suppose you select a specific game *g* and your goal is to show that new model *m* with $\boldsymbol{G}^m$ (potentially different from $\boldsymbol{G}^N$ when *m* has unusual preferences) provides a more accurate description of the way participants engage with game *g* than the neoclassical model m^N which views the game *g* as a representative of $\mathbf{G}^{\mathrm{N}}$. The goal of this chapter is to first convince you that (often) a single game *g* may not be sufficient to reach strong conclusions as to the general validity of your new model m.[36] In this chapter I describe in detail various strategies that help you do that. Finally, I urge you to consider a stress-test for your argument of a new model m, so you and your readers can be even more certain of your conclusions.

3.1 Testing between Two Models: Beware the *Background Noise*

Consider a Game $\boldsymbol{G}^N$ described by the neoclassical model m^N and a Game $\boldsymbol{G}^*$ described by the new or unusual model m^*. The difference of the new model m^* from the neoclassical model m^N could be a different utility function, like adding utility over ego-relevant beliefs, altruism, or competitiveness (in which case $\boldsymbol{G}^*$ is likely different from $\boldsymbol{G}^N$); the presence of a friction, such as costs of computing outcomes, or computational mistakes that arise from complexity (in which case $\boldsymbol{G}^*$ is likely the same as $\boldsymbol{G}^N$ but the constraints to compute solutions and equilibria differ); or an unusual mental model, like the level-k model, or failures in contingent thinking (in which case $\boldsymbol{G}^*$ is likely the same as $\boldsymbol{G}^N$ but the solutions and equilibria differ). Consider a specific *game g* implemented in the experiment, where payments are in money rather than utils. If $(m^N, \boldsymbol{G}^N)$ does describe all the forces present when participants play the *game g*, then we expect participants to behave according to the neoclassical prediction $\hat{a}(m^N,g)$. However, if instead $(m^*, \boldsymbol{G}^*)$ is a better description of the forces guiding behavior in game g, then we expect participants to select actions that conform to the prediction by m^*, namely $\hat{a}(m^*,g)$.[37]

Suppose the experiment delivers actions $\tilde{a}(g)$ that are closer to the predictions using the new model m^* rather than the neoclassical model m^N. An obvious first check, I won't even spend too much time on, is to assure that there is not another equilibrium of the neoclassical model that, perhaps, predicts actions that are even closer to those observed in the data. When this is not the case, it may be very tempting to declare victory, believe that the experiment provides clear evidence that the data are consistent with the new or unusual model m^*.

However, the data we observe in the experiment do not *just* depend on the abstract Game $\boldsymbol{G}^N$ or $\boldsymbol{G}^*$, they may also depend on the specific *game g* implemented in the experiment and how we implemented it, as well as whether there are other forces present beyond just the two currently considered models m^* and m^N. I call the combination of all of these forces "***background noise***" $\boldsymbol{B}$. The main insight is that the data we observe, $\tilde{a}(g)$, which we hope to be influenced either just by m^* or m^N, are in fact *also* influenced by $\boldsymbol{B}$, and as such are more realistically described as $\tilde{a}(\boldsymbol{B},g)$ than $\tilde{a}(g)$.

Before I show how we can assess the role of the *background noise* $\boldsymbol{B}$ in affecting actions of experimental participants, I describe several forces contained in the *background noise* $\boldsymbol{B}$.

[36] For example, suppose in the ultimatum game played for $10, where offers have to be in units of one dollar, you observe player 1 proposing $5 to the other player and player 2 accepting the offer. Note that this is far from the two possible subgame perfect equilibria of proposing either $0 or $1. However, it is a Nash equilibrium. It could also be players follow the prediction because players have a utility function that is best described by m* where the utility of a player is given by the minimum of their own and the other players payoff. While it is tempting to conclude that that the game provides evidence of m* it could be due to the fact that player 2 gets angry when they are offered an amount less than half the total pie, and when player 2 is angry they receive more pleasure from destroying the pie than from accepting any offer. Player 1 knows that, and this is why they offer half the pie, not because player 1's preferences are other-regarding. We cannot distinguish between those two models from just observing behavior in the ultimatum game if player 1 offers half the pie.

[37] For simplicity, throughout this paper, I will focus on actions rather than strategies.

3.1.1Procedural Noise

The *procedural noise* part of the *background noise* ***B*** consists of all kinds of experimental details: One is the exchange rate from payoffs in the experimental currency to benefits to the participants, which varies from small to large stakes, as experiments largely incentivize decisions. Another detail is whether the game or decision is played once (one-shot experiment) or multiple times, and in the latter case, whether all decisions (or all rounds) are paid or just one. Details also include the myriad choices that are made when framing the game *g* via the interface and the particular instructions. Finally, they include idiosyncrasies of the specific participant population with their education, norms and expectations, their strategic sophistication, etc.

There are broadly three different ways in which the experimenter, via experimental details, innocently (and hopefully not nefariously) may affect and manipulate the observed actions. Note that this can look like "magic," and might be part of the problem why economists worry that experimenters can generate environments where "the theory doesn't work." However, I think as empirical economists we "kid ourselves" if we believe that *procedural noise* only plays a role in the laboratory and outside of the lab actions just depend on our stylized version of the abstract Game ***G*** of the – honestly, in general – only partially observed environment. In the field, forces that are the equivalent of procedural noise in the laboratory are given by the specific, special and rich context with its accompanying potential norms and habits, as well as specific idiosyncrasies of the population.

The first of the broadly three ways an experimenter may affect observed actions is that we can make individuals select decisions at random: In the most extreme version, we could provide American undergraduates with instructions and decision screens in German, perhaps in a five point font. Clearly, declaring victory because one fails to find a specific significant effect would be foolish, as we, effectively, just managed to have individuals make random decisions. While this is an extreme example, it may be easier than we expect to generate excess randomness. This could be due to confusing instructions, including confusing ways to describe how choices translate to payments, or presenting participants with problems that are simply too difficult to solve, to name a few examples.[38]

A second way in which a hapless (or nefarious) experimenter may receive "meaningless" results is through floor or ceiling effects. Suppose I would love to receive a result where most individuals, when selecting between the X and Y button, select X, as this would validate my hypothesis. A most extreme way to achieve this is to make the Y button very hard to find, or difficult (or impossible) to press, or result in a costly delay rather than an instant decision. All these are ways to make the Y option more costly and steer individuals into selecting X, even though not necessarily because of my model m^*. Of course, no honest experimenter would deliberately do this, but it could be that involuntarily someone designs an experiment that in essence achieves such an effect.

A third way the experimenter may affect decisions is via the use of instructions or the interface by manipulating how a participant perceives the game. This can be done by emphasizing certain aspects of the game, use instructions that propose a certain behavior, or via the use of specific examples. For example, Binmore, Shaked, and Sutton (1985) found, compared to other studies, very low offers in the ultimatum

[38] For example, suppose participants are selecting between two options, each of which represents a payment that can be calculated using a difficult mathematical expression. If participants cannot solve these problems, we would not expect their choices to reflect their true preference over money.

game. However, in their instructions, they write “How do we want you to play? YOU WILL BE DOING US A FAVOUR IF YOU SIMPLY SET OUT TO MAXIMIZE YOUR WINNINGS.” where the text is exactly as it was in the instructions, font included.[39] A second example is from Liberman, Samuels and Ross (2004) who describe the same prisoner’s dilemma game to participants as either a community game or a wall street game. They show that this affects behavior. These last examples emphasize that when participants play game *g*, instructions, the framing, the description and the depiction of the game *g* may trigger forces we may not have thought of, which brings me to the second component of the background noise.

3.1.2 Other Forces in the Background Noise B

We spent the end of the last chapter talking about *g*-hacking: The fact that a researcher may try out various representations of a game *g* each of which only in theory, but not in practice, deliver “similar” results according to their description as a game in $\boldsymbol{G}^N$ or $\boldsymbol{G}^*$. If for two specific representations *g* and *g*’ the results are, ceteris paribus (that is keeping all else constant), reliably different, despite the fact neither the new nor the neoclassical model predict a change, then there are clearly forces at work when participants play the game *g* which are not described by neither m^* nor m^N.

Sometimes, such other forces that are part of the *background noise* ***B*** may be well understood and have their own transportable models. We return to such “multiple model testing” in Sections 3.3 and 3.4. Sometimes, such other forces are less well understood or easily described in a general way. In Sections 3.1.4 to 3.1.5 I provide specific examples of such *background noise*. While it may be hard to think of all possible forces that may operate on *g* but that are not described by either the new or the neoclassical model, it is also clear that they are present! This is because as experimental and behavioral economists we still discover new forces and models that were previously unknown, but nonetheless presumably were always affecting behavior. Examples of such new models are issues of complexity (which we revisit later), attention (for an overview see Loewenstein and Wojtowicz, 2023) or salience (for an overview see Bordalo, Gennaioli, and Shleifer, 2022) that guide behavior either because of the specific game *g* or the specific experimental implementation with its corresponding *procedural noise*. For a nice experiment that points out how manipulations of attention can affect choices, see Conlon (2024). These forces, that are neither captured by the neoclassical model m^N nor the alternative model m^*, can lead to behavior that is neither random nor on the boundary and as such not necessarily easily detected as being an artefact of the *background noise* ***B***. Put differently, just because there is (reasonable) variance in the data does not imply that the Experimenter found an environment with an implementation that allows to just hone in on m^N and m^* as potential explanations for the behavior.

Note that these forces are present outside of the laboratory as well: There are whole fields dedicated to attention and how to affect individuals’ choices: advertising, human factors and specifically human factors engineering, or the subfield of computer science that deals with data and scientific visualization, and in Economics, the work on behavioral economics, especially the literature on choice architecture. For example, Bertrand et al (2005) consider South African lender who sent letters offering clients large, short-term loans at randomly chosen interest rates. They find that the interest rate significantly affected loan take-up, but so did some of the psychological features added in the letters (these were specific types of frames

[39] At the time this created a large controversy. Personally, I do not see a huge issue in having different and perhaps unusual preferences. This only becomes a problem if the researcher is not open about it and hides this fact.

and cues shown to be powerful in the lab, but which, from a normative perspective, ought to have no impact).

When the combination of forces that are the equivalent of *procedural noise* and *other models* in the field are very strong, they may be responsible for the myriads of projects that do not see the light of day. The same is true in experimental economics, when we do not manage to "tame the ***B***." So, how then can we control for the *background noise* ***B***?

3.1.3 Controlling for the Role of the Background Noise B

Before detailing how to control for the *background noise*, it is important to remember that we can influence its importance, and transparency. As an experimenter, I am often asked why we have such a preference for clean designs. This is exactly why. A clean design that does not introduce superfluous elements is one that reduces potential forces in the *background noise* and makes the remaining ones as transparent as possible.

We start with a game *g* which, given the action space and the monetary payoff consequences of actions, is plausibly a representative of belonging to the Game $\boldsymbol{G}^{N}$ or to the Game $\boldsymbol{G}^{*}$. The neoclassical model m^{N} predicts actions $\hat{a}(m^{N},g)$ while the alternative, new or unusual (relative to m^{N}) model m^{*} predicts different actions $\hat{a}(m^{*},g)$. Assume the experiment delivered actions $\tilde{a}(\boldsymbol{B},g)$ that are closer to the predictions of the new than the neoclassical model, that is closer to $\hat{a}(m^{N},g)$ than to $\hat{a}(m^{*},g)$. Since I provided a myriad ways in which the *background noise* may "move around" actions, the crucial question is: Are the observed data $\tilde{a}(\boldsymbol{B},g)$ closer to $\hat{a}(m^{*},g)$ than $\hat{a}(m^{N},g)$ because m^{*} is indeed a better description or predictor of the data, or because of distortions driven by the *background noise* ***B***?

One way to show that m^{*} is a crucial factor for the observed actions $\tilde{a}(\boldsymbol{B},g)$ is a **design by elimination** (see also Niederle, 2015). Basically, construct a new *game* $g^{Eliminate\ m*}$ such that: (i) The neoclassical prediction of behavior in game $g^{Eliminate\ m*}$ is identical to the prediction in *g*, that is $\hat{a}(m^{N},g)= \hat{a}(m^{N}, g^{Eliminate\ m*})$. (ii) The change in procedures and instructions is minimal so that -- as much as possible, and to the best of our knowledge -- we would expect the *background noise* $\boldsymbol{B}^{Eliminate\ m*}$ to affect actions in $g^{Eliminate\ m*}$ the same way the *background noise* ***B*** affected actions in *g*.

Suppose we are convinced we found a game $g^{Eliminate\ m*}$ that fulfills the two requirements. If we find that $\tilde{a}(\boldsymbol{B}^{Eliminate\ m*}, g^{Eliminate\ m*})$ is different from $\tilde{a}(\boldsymbol{B},g)$, then we have shown that m^{*} is necessary to produce behavior in the original *game g*. The difference between $\tilde{a}(\boldsymbol{B}^{Eliminate\ m*}, g^{Eliminate\ m*})$ and $\tilde{a}(\boldsymbol{B},g)$ serves as a measure of the effect of m^{*}.[40] Hence, behavior in $g^{Eliminate\ m*}$with the *background noise* $\boldsymbol{B}^{Eliminate\ m*}$ serves as a measure of the "but for" world, the "alternative universe" that we would observe in the absence of m^{*}. I hope that after reading so far, I convinced you that it may be problematic to *just* assume that this alternative universe is equal to the neoclassical prediction of behavior in *g* and hence *just* equal to $\hat{a}(m^{N},g)$.

At what point can we be certain that we found such a game $g^{Eliminate\ m*}$ with corresponding *background noise* $\boldsymbol{B}^{Eliminate\ m*}$? The difficulty may not be in finding a game where m^{*} has no "bite," but in being sure the game $g^{Eliminate\ m*}$ is similar enough to the original game *g* that the assumption that the impact of the *background*

[40] Note that, strictly speaking, we identify the effect of m^{*} on the game *g* only when using the specific *background noise* ***B***. It may well be that the *background noise* has a feature or force that is necessary for m^{*} to operate on *g*, even though we "forgot" to put it in the model m^{*}.

noise $\boldsymbol{B}^{Eliminate\ m^*}$ on actions in the new game $g^{Eliminate\ m^*}$ is similar to the impact of the old *background noise* $\boldsymbol{B}$ on the old game *g*.

Note that to prove the role of m^* in generating the observed behavior in game *g*, you really only need to make the case that the expected impact of the new *background noise* $\boldsymbol{B}^{Eliminate\ m^*}$ on actions in the new game $g^{Eliminate\ m^*}$ are such they work *against* proving the role of m^* to account for behavior in the original game *g*. Put differently, suppose the new *background noise* $\boldsymbol{B}^{Eliminate\ m^*}$ affects and distorts actions in game $g^{Eliminate\ m^*}$ differently than the original *background noise* $\boldsymbol{B}$ affected and distorted actions in the original game *g*. If this additional or new distortion is such that the new *background noise* $\boldsymbol{B}^{Eliminate\ m^*}$ moves actions in the new game $g^{Eliminate\ m^*}$ *towards* actions corresponding to $\tilde{a}(\boldsymbol{B},g)$, the observed actions in the original game *g*, then you bias yourself *against* finding that the model m^* was necessary to generate the actions $\tilde{a}(\boldsymbol{B},g)$ you observed in the original game. It may lead you to *underestimate* the role of m^* in generating the observed outcomes in the original game *g*. Note that as a rule or for practical reasons, this is of course not ideal, but it is ok as long as you still have results – that is, observed actions in $g^{Eliminate\ m^*}$ are much closer to the alternative model that the observed actions to $\tilde{a}(\boldsymbol{B},g)$ were in the original game. As a consumer of your findings, we "mostly" worry if there is a chance you made your life too easy, and especially easier than your hypothesis allows for. Making your life harder is not ideal, but it ok.

Suppose we find instead that $\tilde{a}(\boldsymbol{B}^{Eliminate\ m^*}, g^{Eliminate\ m^*})$ and $\tilde{a}(\boldsymbol{B},g)$ are identical. Can we, in turn, affirm that m^* plays no discernable role in the behavior of participants when playing *g*, and hence is a model that has no predictive power? Strictly speaking, we can only infer that the force m^* is not strong enough to affect behavior in *game g* above and beyond the forces in the *background noise* ***B.*** For example, consider the case where the neoclassical model predicts participants to select the X button and the new model m^* predicts participants to select the Y button. Then eliminating the force m^* in a new game $g^{Eliminate\ m^*}$ can only affect behavior if initial Y choices were due to m^*. However, suppose instructions are in five-point font and participants are making random choices. Then using $\tilde{a}(\boldsymbol{B}^{Eliminate\ m^*}, g^{Eliminate\ m^*}) = \tilde{a}(\boldsymbol{B},g)$ to conclude that m^* can never affect behavior in *game g* is erroneous, as this is may be the case only because of the specific overwhelming *background noise* $\boldsymbol{B}$ which turns participants into quite completely unresponsive and random button pressers.

While $g^{Eliminate\ m^*}$ is a game and experimental treatment that shows the alternative universe when m^* is eliminated and has no "bite," you can think of $g^{Eliminate\ m^*}$, alternatively, as a control treatment of the *background noise* $\boldsymbol{B}$. Put differently, it is almost as if instead naming the treatment one where eliminate m^*, we name it the one where we hone in on all other hypotheses and forces, and especially also the forces "hidden" in the *background noise* $\boldsymbol{B}$. As such you could also think of it as a background noise control treatment (in case you otherwise have only one alternative hypothesis, the neoclassical model, which you do not turn off).

3.1.4 Specific Background Noise: Computational Problems

I present two examples from my own work to describe such *background noise* control treatments. In Mobius, Niederle, Niehaus and Rosenblat (2022), we study whether individuals are non-Bayesian when updating beliefs that hold ego-relevance. The motivation was to show evidence of non-Bayesian overconfidence.[41] To study updating and deviations from Bayes rule, we needed an environment where we

[41] At the time we ran the experiment in the early years of this Millennium (we had to describe what "The Facebook" was, as it was then called) such evidence consisted mostly of data that more than X% of participants believe that they

could elicit beliefs in sufficient detail to check whether individuals are Bayesian. We (re)discovered that asking individuals about a binary event, in our case their belief whether they were in the top half of all students taking an IQ-style test, requires only eliciting one number that completely summarizes an individual's beliefs.[42] After eliciting a prior, we gave participants a signal that was correct with 75% chance and elicited a posterior, a process we repeated four times.

We showed that individuals who update about their chance to be among the top half performers update conservatively – that is, acted as if the signal was less informative than it was – and asymmetrically – that is updated more after a positive than a negative signal. Our paper included a model that shows that an individual whose beliefs about their relative IQ performance is ego relevant would update conservatively and asymmetrically, our m^*, so to speak.

Should we have stopped there and declare victory, as the data $\tilde{a}(\boldsymbol{B},g)$ were closer to $\hat{a}(m^*,g)$ than to $\hat{a}(m^N,g)$ where the latter is perfect Bayesian updating? Is the "alternative universe," the "but for" world without distortions due to ego-relevance of beliefs truly best described as Bayesian updating? Put differently, how much of the non-Bayesian updating was due to our model on ego-relevant beliefs providing utility, m^*, and how much to individuals being, for example, just bad Bayesians? That is, how much of the deviations from Bayesian updating are due to general *background noise* $\boldsymbol{B}$? While asymmetry may be hard to justify as a "random mistake," conservatism may well be due to a general problem of individuals' ability to Bayesian update.

We therefore needed a *background noise* control treatment to measure how individuals would update their beliefs in the absence of ego-relevance that is otherwise as similar as possible, that is we created an environment where we eliminated forces in m^*. Instead of updating about the potentially ego-relevant event to be among the top half of performers in an IQ test, participants updated beliefs about a random event. With instructions being identical (apart from the description of the binary event on which participants received signals), we felt it safe to assume that the impact of the *background noise* $\boldsymbol{B}^{Eliminate\ m^*}$ on behavior in this new game $g^{Eliminate\ m^*}$ was virtually identical to the effect of the *background noise* $\boldsymbol{B}$ on ego-relevant updating in game g. Furthermore, the neoclassical prediction in games $g^{Eliminate\ m^*}$ and g is identical, namely Bayesian updating, hence $\hat{a}(m^N,g)= \hat{a}(m^N, g^{Eliminate\ m^*})$.[43] While participants were still not Bayesian, they were significantly less conservative and not asymmetric. Hence, the new treatment allowed us to confirm that while our experimental participants were not perfect Bayesians when updating beliefs about non-ego-relevant events, their deviations from Bayesian updating were larger (more conservative) and biased (asymmetric) when their beliefs were ego-relevant.

The second example is from Martinez-Marquina, Niederle and Vespa (2019), where we used two different "alternate universes" to ensure that results in the main treatment are not driven by *background noise*. Recall from Section 2.2 that participants submit a price p to buy a firm that has one of two equally likely possible

are in the top X percent. However, theory tells us that this, on its own, while perhaps surprising, is not evidence of any non-Bayesian updating, as all depends on the data generating process. For example, suppose half of all drivers are good drivers and the others are bad drivers. A bad driver is someone who has an accident in the first ten years of driving. Then, after five years (and a few realized accidents), more than fifty percent of people will correctly believe that they are more likely to be good than bad drivers, see also Benoit and Dubra (2011) and Moore and Healy (2008).

[42] Note that beliefs about the relative or absolute performance are complicated distributions. But the belief whether the performance is in the top 50% is just one number between 0 and 1, the probability that this is true.

[43] We even ensured that the priors in g and $g^{Eliminate\ m^*}$ were very similar by providing individuals with an objective prior in $g^{Eliminate\ m^*}$ that was within five percentage points of their prior in g.

values, $v_L=20$ or $v_H=120$. A participant who submits a price p to a firm of value v makes profits $1.5v-p$ if $p \geq v$ and 0 otherwise. The well-known result in the literature, that we replicate in our experiment, is that many participants submit a price of $p=120$ and hence incur expected losses. We further show that these same participants prefer a 50:50 lottery of (10, 0) – the lottery resulting from $p=20$ -- to one of (60, -90) – the lottery resulting from $p=120$, and hence are not sufficiently risk-seeking to make $p=120$ payoff maximizing.

Furthermore, we explicitly show, rather than just assume, that when participants submit a price for a firm of one known value, they are very likely to submit $p=v$, the payoff maximizing price (this is the first "alternate universe" we explicitly measure). What then makes the problem when the value of the firm is one of two possible values so much more difficult? The main message of our paper is to point out that there are two changes that occur: One is moving from one state or contingency (or firm-value) to two, and the other is moving from realized, certain contingencies to potential or hypothetical ones. To disentangle the potential role of computational complexity from the role of risk, we invented the complexity control treatment: What are the problems of participants when they must consider two states, that is, what is the role of the computational complexity? And is there some added source of mistakes when the participant moves from a deterministic environment with two states to one that in addition involves risk?

We therefore constructed a control for the complexity of the computations involved to account for "deviations" from expected profit maximizing choices in the *absence* of hypothetical states: In $g^{Computations}$ a participant submits a price p which is then transmitted to both the low value firm $v_L=20$ and the high value firm $v_H=120$. Hence, the participant, dependent on the submitted price p, purchases no firm, just the $v=20$ firm or both the $v=20$ and the $v=120$ firm. Note that for each p the certain, deterministic payoffs in this computational complexity only treatment are identical to (twice) the expected payoffs in the main treatment. In our paper, we named the two treatments the deterministic and probabilistic treatment, respectively.[44]

For a participant who is not risk-seeking, $\hat{a}(m^N,g)= \hat{a}(m^N, g^{Computations})$ namely $p=20$. Furthermore, we wrote instructions and made decision screens for those two treatments simultaneously to ensure minimal differences. In both treatments, a participant had to consider both firms to compute the payoff maximizing price. The difference is that in the original game g the contingencies, states or realized values of the firm were hypothetical, while they are realized in $g^{Computations}$. We find that, comparing behavior in g and $g^{Computations}$, participants were significantly more likely to submit the (expected) payoff-maximizing price in the deterministic, the complexity only treatment, than in the main probabilistic treatment.

Overall, we find that, compared to the case of only one firm, participants are less likely to submit the payoff maximizing price when they have to consider two firms without any uncertainty. However, as soon as risk is involved, because firms are hypothetical or potential rather than realized firms, the problems of participants are significantly magnified: In our experiment, in the probabilistic treatment, only about half of the failures to maximize payoffs are due to computational complexity, while the other half is due making states uncertain, or hypothetical.

We think that the effects of the *background noise* $\boldsymbol{B}^{Computations}$ on actions in $g^{Computations}$ are similar to those of the baseline *background noise* $\boldsymbol{B}$ on actions in the baseline (probabilistic) game g. Nonetheless, we separately stress-tested the conclusion that the behavior in the complexity control treatment was indeed

[44] In the probabilistic treatment, we multiplied payoffs by two to equalize expected payoffs for each submitted price p.

driven by the participant submitting a payoff-maximizing price rather than any other possible new reason, some new unknown model that may suddenly rear its head in $g^{Computations}$ but not in g. I discuss this in Section 3.5. We also stress-tested the insight that additional problems arise when participants make decisions in the standard, probabilistic treatment than in the deterministic complexity control treatment. I will come back to the importance of such stress-tests in Section 3.6.[45]

Without the computational complexity control treatment, we might have misattributed some of the non-payoff maximizing choices to unusual preferences, such as wishful thinking in Martinez-Marquina, Niederle and Vespa (2019) or risk-seeking preferences in Martinez-Marquina, Niederle and Vespa (2019) and Niederle and Vespa (2019).

The computational complexity control treatments in Martinez-Marquina, Niederle and Vespa (2019) and Niederle and Vespa (2019) point out that not all (though some) non-expected payoff maximizing choices are mistakes present even in the absence of risk, when participants only have to compute payoffs. While work by Oprea (2024) suggests an even larger role for computational complexities in accounting for unusual choices, all these papers point to computational complexity being a factor in decisions under risk. It remains important to address to what extent unusual decisions in choices under risk are due to complexity, see also Enke and Graeber (2023) for an influential contribution to this literature and Niederle and Vespa (2023) and Enke (2024) for overviews on the rising literature in economics on the cognitive approach.

3.1.5 Examples: Background Noise Controls Appeared Later

The examples of background noise control treatments I presented, including the specific "computational complexity only" control for choices under risk in Martinez-Marquina, Niederle and Vespa (2019) and Niederle and Vespa (2019) (termed mirror by Oprea, 2024), beyond being from my own work, are all from experiments where the original paper included a treatment to understand to what extent results are driven by the *background noise*. This may lead you to the unfortunately erroneous view that clearly no experiment or known economic result would suffer from a contamination of *background noise*. I will provide two examples from experimenters I admire, where a paper containing a control for the *background noise* was done only much later than the initial paper. Both the prominence of the results, and the fact that the papers measuring the *background noise* came so much later, shows that in the profession as a whole, not just among

[45] After inventing the computational complexity control treatment, the deterministic treatment, for the "acquiring-a-company" problem in Martinez-Marquina, Niederle and Vespa (2019), we applied this methodology to choices over lotteries, see Niederle and Vespa (2019). When considering a participant who selects one of two lotteries, then, in many instances, either choice can be rationalized by a specific risk parameter, or perhaps by individuals being risk-seeking. We wanted to know how many (or all?) choices that deviate from risk-neutrality are driven by preferences or by mistakes due to computational complexity. To turn this around, we asked whether many choices that suggest that the participant deviates from risk-neutrality arise even in a slightly modified (the deterministic) environment where the only problem is the complexity of computing payoffs and where risk preferences play no role (that is there is a dominant choice)? We therefore had participants not only select among two lotteries, but also introduced a complexity-only control treatment, where participants selected among the respective certain payoffs, the expected value, where options were described similarly to the lotteries. Such a control for computational complexity is also used in Oprea (2024), who introduced the term mirror for the complexity-only control treatment. While we find that (the few) choices that suggest risk-seeking preferences are likely due to computational complexity problems, participants do seem to be risk-averse to an extent not mirrored by choices in the deterministic counterpart. Our results are hence quite different from Oprea (2024) who finds that computational complexity on its own yields results virtually indistinguishable from decisions in the lottery domain.

specific authors, it is not always easy to see the need nor how to control for *background noise*.[46] The first example had a paper containing the *background noise* control decades ago, while the corresponding *background noise* control for the second was published just now. In both instances the role of the original hypothesis or force was not eliminated, but the lower bound of its effect was significantly lowered. I end with an example where the evidence attributed to the hypotheses in a specific design was entirely by *background noise.*

The notion of altruism and specifically that individuals may not fully free-ride when it comes to contributions to public goods, was one that was very controversial in Economics, see e.g. the early paper by prominent sociologists Marwell and Ames in 1981: "Economists free ride: Does anyone else? Experiments on the provision of public goods."[47]

After several papers on public goods, including an early paper by Jim Andreoni (1988) himself, he wrote a paper in 1995 called "Cooperation in Public-Goods Experiments: Kindness or Confusion?" He writes in the abstract (page 891) that "[t]his paper presents the first systematic attempt to separate the hypothesis that cooperation is due to kindness, altruism, or warm-glow from the hypothesis that cooperation is simply the result of errors or confusion […]."

In each of the ten rounds, 20 participants are randomly matched into groups of 5. The 5 participants each receive 60 points they can keep or decide to contribute some (or all) to the public good. The public good equals to the sum of the five individual contributions and each participant receives 0.5 points for every point in the public good. In the regular public good game treatment, participants are paid linearly as a function of the points they made in each round. To control for whether donations are driven by motives such as altruism or rather errors and confusion, Andreoni constructs a "eliminate altruism" treatment (or if you wish, *background noise* control treatment), where he eliminates any motives pertaining to altruism (or kindness or warm-glow), while only introducing minimal changes.

In this new treatment participants play the same game, but now, instead of paying participants linearly for their points in each round, participants are paid based on the ranking of their points. Specifically, the participant with the most points in a given round receives a fixed amount, the one with the second highest points a lower fixed amount, etc, that is, basically participants now play in each round a zero-sum game of who has more points.[48] Converting points to payments via rank rather than linearly does preserve the dominant strategy equilibrium of not contributing to the public good. However, payments through rank eliminate any incentives for cooperation.[49] Andreoni (1995) finds a significant amount of confusion, that is positive donations even in the eliminate altruism treatment (or *background noise* control treatment).

[46] Note that experiments are also very much creatures of their time, as alternative hypotheses for the behavior that needs to be controlled and accounted for varies over time.

[47] Ledyard (1995) in his review on public goods notes that "[m]ost economists believed there was a free rider problem and that voluntary contribution mechanisms would provide very little public goods." (p 122).

[48] In the paper, Andreoni (1995) notes that a second change was made in the eliminate altruism treatment compared to the regular treatment: Participants paid based on their rank learned about their rank in each round. It could be that any change in behavior may be due to receiving information on relative earnings rather than the elimination of other-regarding motives. Andreoni therefore included a third treatment where he added the rank information to the regular public good game. One way to think of this is that he modified the main treatment to even further reduce the differences between the game that includes and the one that excludes other-regarding concerns.

[49] Specifically, a participant, who, say, values others' payoffs at some fraction (less than 1) may contribute to the public good when points translate to payoffs linearly and contributing to the public good increases total payoffs, but not when points translate to payoffs via rank.

However, contributions do not reach the levels observed in versions of the game where cooperation increases total payoffs. In general, experiments that involve payoff-maximizing choices that are on the boundary and where the new model m^* predicts interior choices will always suffer from confounding choices driven by m^* with choices driven by confusion or noise. For a more recent systematic evaluation of the role of noise in generating deviations from the Nash equilibrium towards Pareto (total payoff maximizing) outcomes, see Recalde, Riedl and Vesterlund (2018). They avoid boundary concerns by using public good games where both the Nash and the Pareto outcome are in the interior.

The second example is the moral wiggle room, the idea that individuals avoid information to maintain positive views about themselves, even when this avoidance may lead to actions that suggest the opposite. In the seminal paper by Dana, Weber and Kuang (2007), participants select between two options, A and B. They know that their earnings are larger from taking action A, but they do not know which action is relatively harmful and which is relatively beneficial to another participant. The participants have the choice to either select the action immediately, or to first find out which of the two underlying potential payoffs for the other participant are realized when selecting action A.

The robust and often replicated result is that many individuals opt to avoid information and select action A, even though those same individuals often do not select A when they know that it is relatively harmful to the other participant. That is, when payoffs are known, participants are often willing to sacrifice a small amount to help the other, but when payoffs to the other are hidden, participants prefer to avoid payoff information and select the option that is best for themselves. As such the paper provided evidence that concerns about others' welfare are not purely a reflection of a preference for social welfare. Rather, Dana, Weber and Kuang (2007) conclude that some fairness behavior may mainly be due to the fact that many participants intrinsically dislike appearing unfair, either to themselves or others.

Exley and Kessler (2023) point out that explanations of image concerns "…however, rely on the sophistication of agents to strategically avoid information in order to maintain certain beliefs or in order to construct plausible deniability about their actions." It could be that the original result might be due to *background noise*: Is it really the case that, absent image concerns, agents would select to acquire information about the impact of each action on the other participant? To check whether participants strategically avoid information because of image concerns (rather than other reasons), they construct a new treatment with a new game $g^{NoImage}$ that is as close as possible to the original game g with very close experimental procedures and instructions, hence the impact of the new *background noise* $\boldsymbol{B}^{NoImage}$ on actions being as close as possible to that of the original *background noise* $\boldsymbol{B}$. However, actions in $g^{NoImage}$ are not driven by image concerns, that is the main hypothesis, image concerns, has no bite in $g^{NoImage}$. In $g^{NoImage}$, the decision maker does not receive any payoffs, rather they go to a third participant. That is, the participant, instead of making decisions that impact herself and another, decides for a third participant, where decisions impact that third participant and another, but not the decision maker herself. As a result, this third party treatment removes any motives of image concern for the decision maker. First, note that worries to appear selfish play no role as there is no room for selfishness. Second, a desire to appear fair would, if anything, lead the decision maker to be more, rather than less likely to uncover payoff information before taking an action. Put simply, compared to the standard treatment, in the $g^{NoImage}$ treatment acquiring information does not lead to a trade-off between a choice motivated by image concerns and a choice motivated by selfishness, as, by design, there is no room for selfishness.

Exley and Kessler (2024) find that many participants fail to uncover information in this third-party treatment $g^{NoImage}$. However, even more participants opt to avoid information when image concerns are introduced. That is, image concerns play a role, just not as large one as we initially thought. To summarize, they write what is necessary for a *background noise* treatment: "We compare the rates of (passive) information avoidance in the classic Dana, Weber and Kuang (2007) setting to a new setting that makes minimal changes to remove image motives to avoid information; our new setting holds constant the structure of the decision, the content of the information, and the timing of information provision. We then attribute to image concerns any difference in information avoidance—by which we mean subjects failing to acquire easily accessible information—across the two settings."

3.1.6 Examples: Background Noise Generated the Result

I provide one example, where a failure to account for background noise led to non-robust results.[50] Rand, Greene and Nowak (2012) introduced the Social Heuristics Hypothesis, which argues that fast, intuitive responses are shaped by past experiences in repeated interactions where cooperative behavior may be optimal. That is, for a given situation, fast responses are predicted to lead to more pro-sociality than slow responses. Rand, Greene and Nowak (2012) consider several economic games and find that participants who reach their decisions more quickly are more cooperative. Acknowledging that a correlation is not equal to causation, they also have treatments where they force participants to decide quickly, which increases contributions, while forcing them to decide slowly decreases contributions.

Several papers could not replicate the result, which resulted in a Registered Replication Report:[51] Bouwmeester et al (2017) find that the increase in contributions when participants are forced to decide quickly is due to selection. Specifically, consider choices where participants have a long time to make a decision. Now force participants to decide quickly. Some participants fail to reach a decision quickly and have their choice eliminated, this alone leads to change in contributions.

To circumvent selection problems, in Kessler, Kivimaki, Litwin and Niederle (2024), we use a method – adopted from work on rational inattention (Caplin, Dean and Martin, 2011) – that allows us to observe incentivized choices of the same individual over the course of a minute. Participants play 10 prisoner's dilemmas and 10 dictator games. In each game, participants decide whether to transfer \$1 to give \$X to another participant, where the exchange rate \$X (i.e., the efficiency of giving), ranges from \$0 to \$10 across games. In our main treatments, participants have 60 seconds during which they can record an initial answer and then subsequently change their answer if they would like to do so. For the specific game that is chosen for payment, one of the 60 seconds is randomly selected, and the choice recorded by the participant at that second is the choice implemented for payment. If no choice is made at the randomly chosen second, both the decision-maker and the other participant earn nothing. The participant is therefore incentivized to make an (intuitive) initial choice as soon as they have one and to change to a subsequent (deliberate) choice whenever they consider a different choice to be optimal.

[50] For an example where a lot of participants are misclassified due to *background noise*, see McGranaghan, Nielsen, O'Donoghue, Somerville, and Sprenger (2024), the set of authors we discussed already in Section 2.8.

[51] A Registered Replication Report consists of a collection of independently conducted, direct replications of an original study, all of which follow a shared, predetermined protocol. The collection of replications will be published as a single article in *Advances in Methods and Practices in Psychological Science*, and all researchers contributing replications will be listed as authors. For the exact rules and procedures to initiate such a report and what the final publication includes see the website of the Association for Psychological Science.

We find that most participants change their choices over the minute in at least some of the decisions they face. Our main result is that subjects' choices reflect an increasing value placed on social efficiency as they deliberate.[52] Many robustness tests and control treatments serve to determine that our data indeed suggest that participants reliably change their mind, responding more to social efficiency over time when deciding whether to be generous.

We therefore provide direct evidence that decisions involving generosity predictably change when comparing intuitive choices to deliberate ones.[53] That is, we do find that while the original method was very flawed, we are able to find a different method using completely different control treatments, to study the hypothesis at hand.

3.2 Old School Comparative Static Experiments

A significant portion of early experiments, described in Section 3.2.1, were about providing evidence against the neoclassical model rather than providing evidence for a specific alternative model. Subsequently, researchers tested whether the comparative static of behavior across games violates the change predicted by the neoclassical model, see Section 3.2.2 More problematic is using comparative static arguments between two quite different games to test between two specific models (rather than neoclassical or not). The reason is that implicit assumptions necessary for such a test are potentially quite strong, as I show in Section 3.2.3.

3.2.1 Evidence against the Neoclassical Model

In the early days of the behavioral literature, essentially "one-treatment" experiments were somewhat common. The aim was to provide evidence of behavior that is not in accordance with the neoclassical model, demonstrate so-called "anomalies." As such these experiments provided often necessary data to help formulate alternative theories. Famous examples, which I will revisit later, concern choices in risky domains such as the Allais paradox, some of the work by Kahneman, Thaler and Tversky, and early other regarding preferences experiments. In more complex settings, such as auctions, early experiments aim to show that individuals do not bid according to the Bayes Nash equilibrium in private-value first price auctions and suffer from the winners' curse in common value auctions.

Those experiments were not particularly concerned by the extent with which behavior that does not follow the neoclassical prediction is due to *background noise*. *Background noise*, while perhaps not a feature, wasn't necessarily a bug either, as forces that make up the *background noise* at the time may be part of the reason the neoclassical model fails, and many of those forces subsequently received their own models, like other-regarding preferences. It all depends on how "strict" we take the neoclassical model. As a discipline,

[52] That is, when X is low (i.e., when it is more expensive to turn your dollar into a dollar for another subject), subjects on average become less generous over the minute. In contrast, when X is high (i.e., when it is cheaper to turn your dollar into a dollar for another subject) subjects become more generous over the minute.

[53] Our results imply that welfare considerations based on revealed preference may be complicated when analyzing decisions – particularly those involving generosity – that are made without deliberation. Our results also suggest how to encourage individuals to privately provide public goods and to donate to charities. The findings of our experiments suggest that less efficient charities should encourage fast donation decisions (e.g., by asking for donations in time-sensitive situations like when checking out at a store) while efficient charities should encourage potential donors to deliberate about giving.

we are not very good in formulating "the model is mostly right, even though, at the same, of course, in its precise prediction, not correct."

3.2.2Old Comparative Static: Neoclassical or not

When results from such early experiments are robust and become prominent, as Economists we are quite good at coming up with neoclassical reasons why the unusual finding may not be so unusual to begin with. For example, to account for overbidding in first price private value auctions compared to the risk-neutral Bayes-Nash equilibrium (RNBNE), claims were made that this may be largely due to risk preferences, as indeed risk-aversion would result in bids higher than the RNBNE.[54] There were several attempts to understand how much risk-aversion is needed to account for observed bids and discussions whether those are plausible levels or not, see also Bajari and Hortacsu (2005) who take this approach even further, by allowing for individual idiosyncrasies.

Instead of estimating risk preferences, or what risk preferences would reconcile observed bids with the neoclassical model, Kagel and Levin (1993) use a comparative static approach: They exploit that the neoclassical model does not just predict behavior in a given Game ***G***, but in all possible Games, and, as such also how behavior should change as we change the Game. They study three different private-value auctions, the first, second and third price auction, where in each case the highest bidder wins, and the bidder pays either the price attached to the highest, the second highest or the third highest bid, respectively. If *only* risk-aversion accounts for deviations from the RNBNE when bidding in auctions, then, changing the auction format changes the impact of risk-aversion: In contrast to the first price auction, risk-aversion does not affect bidding in second price auctions, that is bids should be in line with participants' private values. Finally, in third-price auctions, participants are expected to bid *above* their private value, and so risk-aversion would lead participants to bid *less* than the RNBNE (rather than more as in the first-price auction).

Kagel and Levin (1993) find evidence against the hypothesis that deviations from bids that correspond to the RNBNE are only due to risk-aversion, especially in larger groups. This suggests that other forces are at play when it comes bidding. Specifically, while they replicate that bids in first price auctions are above the RNBNE (and hence in the direction predicted by risk-aversion), they find that in third price auctions participants in groups of 10 largely bid *above* the RNBNE, the direction opposite to the one predicted by risk-aversion! Since the hypothesis is that deviations from the RNBNE are *only* due to risk-aversion, which can account for overbidding in first price auctions, the experiment on not only first, but also second and especially third price auctions provides compelling evidence that other forces are at play when bidding in auctions, forces that lead to bids above the RNBNE. Furthermore, in at least one of these auctions (the third price auction) these forces are stronger than risk-preferences. In environments in which these unknown forces move in the same direction as risk-preferences like in first-price auctions, it seems then at least dubious that *only* risk preferences are responsible for overbidding. Finally, the evidence clearly refutes that risk preferences account for deviations from the RNBNE in *all* private value auctions.

[54] Another interesting literature (and perhaps somewhat heated exchange) that came out of this debate is the paper by Harrison (1989) on the flat maximum critique. The idea is that overbidding in first-price auctions is not very costly and as such we should perhaps not take the *exact* bids that seriously. In light of using this to explain overbidding in first-price auctions, his paper generated four comments and a reply in the December of 1992 issue of the *American Economic Review*, that I can only highly recommend to read.

To alleviate concerns on the credibility of the assumption that changes in the background noise are not driving the results, one can employ stress-tests, see also Section 3.6. For example, Kagel and Levin (1993) are probably aware of the potential confound that behavior across auction formats may be affected by the different *background noises*, for example due to different instructions (even if that, perhaps on its own, suggests that forces beyond the neoclassical model are at work). They stress-test whether behavior in their auctions follow other predictions from the neoclassical model, beyond whether bids are below, around or above the private value. Specifically, they check whether bids increase, are linear, or decrease in the private value, a prediction of the neoclassical model that may be present even if additional forces guide bidding, such as, for example, the joy of winning. Indeed, they confirm this prediction in all three auctions. This suggests that bids are not unduly distorted by the different background noises in the three games, as bids are neither random, nor do they suffer strongly from floor or ceiling effects.

I am still a huge fan of this paper. Nonetheless, if the goal were to understand why bidders bid as they do, then, in this day and age, I would prefer somewhat closer comparisons across games. This is because there are many changes when we move from a first, to a second or third price auction, and perhaps some of these other changes affect bidding as well. However, these were earlier days, and recall that the goal of Kagel and Levin (1993) was to test whether bids that deviate from the RNBNE are driven solely by risk-preferences.

3.2.3 Old Comparative Static: Neoclassical!

Kagel and Levine (1993) used a comparative static experiment, with quite substantial changes across the two games or environments they study, to refute the hypothesis that *only* the neoclassical model affects actions. Comparative static experiments are also sometimes used to validate the neoclassical model, by showing that the comparative static prediction of the neoclassical model across two different games or environments holds. In this case, the alternative model to be rejected is sometimes one of "no change" (or some simple alternative model whose predictions differ from the neoclassical model).

Such a test often involves two Games ***G*** and ***G'*** that are quite different from one another. The idea is to test whether changes in behavior between games *g* and *g'* representing ***G*** and ***G'***, respectively, follow the comparative static that is given by the predictions of the neoclassical model in each game. However, due to the potentially significant differences in structure between games *g* and *g'*, it is not clear that the implicit assumption for such a conclusion is fulfilled. The implicit assumption (or a closely related one, see Section 3.1.3) is that the effect of the *background noise* ***B*** on actions in g is similar to that of the *background noise* ***B'*** on actions in *g'*. Otherwise, the comparative static of behavior across *g* and *g'* may be driven by differences in the effects of the *background noise* rather than by differences due to the games being different. An additional potential problem arises when either *g* or *g'* have more than one equilibrium. This is especially problematic when it implies that any possible comparative static in behavior can be rationalized.

While this Section may read as if I am not a fan of comparative static experiments, this is not the case, see also Section 3.5. However, such experiments *are* problematic when we want to test a specific hypothesis while, at the same time, not being mindful of the implicit assumption that the effect of the background noise may be different when we consider games that are quite different from one another.

3.3 Multiple Models and Indirect Controls

In this chapter I discuss how to design your experiment when there are multiple potential models or forces you want to address. This becomes relevant when one model, say the neoclassical model, can be adapted to include, for instance, risk preferences, or computational problems, and such a simple adaptation then competes with your new model. In 3.3.1 I describe an ideal design option, by expanding the idea of the design by elimination, the counterfactual when your new force or model m^* has no bite. When a design by elimination is not available, I discuss in 3.3.2 one popular option, namely the use of an indirect control: Measure a potential "nuisance" force separately and then econometrically control for it. In Section 3.3.3 I provide a specific example that highlights the assumptions made during such a measurement exercise.

3.3.1 Design by Multiple Elimination

In Section 3.1 we showed one way to provide evidence that actions resulting from play in *game g* representing the Game $\boldsymbol{G}$ that are in line with a new model m^* are due to this new or unusual model rather than the neoclassical model m^N: the design by elimination. This involves creating a new Game $\boldsymbol{G}^{not\ m^*}$ which has the same predictions under the neoclassical model than $\boldsymbol{G}$ such that the game $g^{not\ m^*}$ played by participants has a similar *background noise* than the game g, but in which m^* has no predictive power. This $\boldsymbol{G}^{not\ m^*}$ is essentially a way to control for the effects of *background noise* $\boldsymbol{B}$ and the neoclassical forces m^N in $\boldsymbol{G}$: In a world without m^*, what would actions in a game $g^{not\ m^*}$ that is as much as possible similar to a game g in $\boldsymbol{G}$ with *background noise* $\boldsymbol{B}$ look like?

This method can also be applied if we aim to provide evidence consistent with m^* but neither with m^N nor some other alternate models m^{A1} and m^{A2} (or $m^{A1}, m^{A2}, \dots, m^{An}$). Now, the trick is to find a $\boldsymbol{G}^{not\ m^*}$ where *all* models but m^* make the same prediction as in $\boldsymbol{G}$, and where, of course, the *background noise* of the implemented game $g^{not\ m^*}$ is similar to the *background noise* of the original game g as a representative of $\boldsymbol{G}$. However, it could be that such a single "alternative universe without m^* but with all other forces including the *background noise*" is hard to find.

In that case, it may be possible to find two alternate universes $\boldsymbol{G}(not\ m^*,\ not\ m^{A1})$ and $\boldsymbol{G}(not\ m^*,\ not\ m^{A2})$, and corresponding games $g(not\ m^*,\ not\ m^{A1})$ and $g(not\ m^*,\ not\ m^{A2})$, each with their own *background noise* that is similar to the *background noise* $\boldsymbol{B}$ in game g as a representative of $\boldsymbol{G}$. Note for this trick to work, we assume that models m^N, m^{A1} and m^{A2} do not interact.[55] If only in g, but in none of the control treatments $g(not\ m^*,\ not\ m^{A1})$ and $g(not\ m^*,\ not\ m^{A2})$ the behavior of participants is in line with m^*, then we have, to the best of our current knowledge, eliminated all other possible alternative models, as well as the *background noise* $\boldsymbol{B}$ as a reason for the unique behavior found in the original game g. That is, we have strong evidence that m^* is a necessary driver to observe the unique behavior we found in game g (with the additional crucial assumption that these various other models do not interact).

For example, in Gneezy, Niederle and Rustichini (2003) we showed that women and men performed similarly in a task when they were paid via a piece rate. In contrast, when they were in a mixed tournament where only the highest performer received a linear payment, we found a significant increase in the gender gap in performance. One possible explanation is that women do not perform highly in a competition against men. But there are many others: Maybe women do not perform highly when they are not sure that they will be paid, perhaps due to higher risk aversion – since in the tournament they have to perform without knowing whether they will be compensated for their performance. However, it could also be that women do not

[55] Specifically, we assume that they do not generate an effect only when combined in a specific way.

respond to tournament incentives or any incentives, perhaps because they are unable to perform higher – indeed, while men increased their performance in the tournament compared to the one in the piece rate payment scheme, women on average did not (note this was an across-subjects experiment). Alternatively, maybe women are more altruistic, and by not increasing their performance aim to collude rather than work a lot just to increase the chance of winning.

While there are more alternative hypotheses, I will focus on these ones for now, for more detail see Gneezy, Niederle and Rustichini (2003). To put these hypotheses into our framework, m^* is that women do not perform highly when competing against men; $m^{W:Risk}$ is that women are, in principle, as responsive to incentive schemes as men are, but women are more risk averse and as such respond differently than men do whenever the payment per correctly solved problem is not certain; and $m^{W:Limited}$ is that women cannot perform higher than their piece rate performance (alternatively, this could also be that women are more altruistic than men are, and hence do not want to perform higher when their performance has negative externalities).

In a different treatment, we had women (and men) perform in single sex tournaments. Since m^* is that "women do not perform highly when competing against men," the single sex tournament has all the forces we discussed that are present in the mixed sex tournament, that is, specifically, $m^{W:Risk}$ and $m^{W:Limited}$, but not m^*. Hence, the single sex tournament is a game representing the Game $\boldsymbol{G}(m^{W:Risk}, m^{W:Limited}, \textit{not } m^*)$. We found that women in single-sex tournaments significantly increase their performance compared to performances under non-competitive incentive schemes. That is, we did not find evidence that women, in general, cannot perform higher, or do not react to tournament incentive schemes. Furthermore, the gender gap in performance in single-sex tournaments mirrors the one in non-competitive payment schemes and is significantly smaller than the gender gap in mixed sex tournaments. That is, while there might be gender differences in risk preferences (though see the discussion in the next subsubsection), they, on their own, do not generate a large gender gap in performance. Put differently, it is only in mixed-sex tournaments that we found a large and significantly larger gender gap in performance than in non-competitive payment schemes or in single sex tournaments, an alternative universe where we take out the mixed-sex composition while keeping the tournament scheme.

While it may appear perhaps a little magical that we found this single-sex treatment, this alternative universe that so cleanly honed in on the mixed sex tournament hypothesis, this is because in our paper we had additional treatments that tried to address just a single concern. Furthermore, it could very well have been the case that the "true" model is that women, while maybe performing highly under increased incentives, don't do so when there is a tournament involved. In that case, the single sex tournament would not have helped us addressing whether $m^{W:Limited}$ is a possible explanation. It is only because results changed in the single sex tournament, the game representing $\boldsymbol{G}(m^{W:Risk}, m^{W:Limited}, \textit{not } m^*)$, compared to the mixed sex tournament, that we were able to draw this conclusion.

Such an experiment where we have several different treatments to address as much as possible various concerns is an experiment that lends itself to a "tree of designs," where basically we learn from the results of one treatment which alternative hypotheses we have not addressed yet, and hence still need to be handled. This allows us to not have to do a whole "tree of designs" ex ante and learn from the results before designing the remaining treatment that addresses remaining concerns. Such an experimental approach has the considerable advantage that we do not have to think of everything in advance, very much in contrast to an experiment that uses a within subject design.

3.3.2Indirect Control

However, sometimes a design by elimination may not be practical, obvious or not feasible (it may just be hard to find such a family of Games ***G'*** where we eliminate the effect of various models and our main hypothesis m^*). Consider bidding in first price common-value auctions, where bids are in general above the risk-neutral Bayes Nash equilibrium, but to an extent that many participants consistently incur expected losses rather than expected gains. This immediately implies that risk preferences alone won't be able to account for such bidding behavior. In Nagel, Niederle and Vespa (2025) we aim to decompose the winners' curse, where we consider three classes of possible reasons why the bids of individuals may deviate from the risk neutral symmetric Bayes Nash equilibrium. These are, as we discussed previously, unusual preferences, computational constraints or frictions, and unusual mental models.

Our (or definitely my) beliefs were that a major driver for overbidding is the failure to condition on winning the auction which, in symmetric equilibria, is equivalent to conditioning on having the highest signal (see Nagel, Niederle and Vespa, 2025).[56] Furthermore, Nagel, Niederle and Vespa (2025) point out that another possible problem may be that participants are not able to compute the expected value of the item correctly, even in the absence of strategic difficulties, that is, even if they condition on having the highest signal. Specifically, given that participants are often bad Bayesians, they may overestimate this expected value. While this is perhaps an obvious explanation, it turns out that it has not received attention in the literature so far. How much, if anything, of overbidding in first price common-value auctions is driven by the failure to compute the expected value of the item conditional on having the highest signal? Generating environments in which participants do not have to bid in an auction, but where those various forces from risk preferences, computational problems or failures of contingent thinking play a role might be difficult. What other strategies do we have at hand?

A popular option to address the role of such alternative hypotheses, such as risk preferences, or not correctly computing the expected value of an item conditional on having the highest signal, is by measuring or estimating them separately. With those measurements in hand, one can aim to econometrically control for them to find out their impact on generating bids that do not conform with the symmetric risk neutral Bayes Nash equilibrium. I call such an approach an *indirect control* for such a force or model.

In terms of measuring risk-preferences, a popular tool is to use one of the standard risk "tasks" like Holt and Laury (2002) or Eckel and Grossman (2002), or some other form of risk-elicitation. For me, there are two problems with such an approach. First, while many behavioral Economists acknowledge that risk preferences are "complicated" and perhaps not always best described by a single parameter, when it comes to controlling for risk, however, often economists (sometimes even the same) are often all too happy to ignore such complications. This may be especially worrisome when gender is one of the studied traits, since different risk elicitations consistently deliver different sizes in the gender gap of risk preferences, ranging from almost no gap (for the Bomb Risk Elicitation task, see Crosetto and Filippin, 2016) to a small gap in the Holt-Laury task (Holt and Laury, 2002) and a large gap in the Eckel-Grossman (Eckel and Grossman, 2002) as well as the Gneezy-Potters task (Gneezy and Potters, 1997), see Crosetto and Filippin (2016) and

[56] Such a failure to condition on the event of having the highest signal can either be due to beliefs of the participant that the bid function of others are such that they may win even if they do not have the highest signal, as in level-k thinking (Crawford and Iriberri, 2007) or cursed equilibrium (Eyster and Rabin, 2005), or because individuals have problems with contingent thinking, see the survey in Niederle and Vespa (2023).

for a survey Niederle (2016). This alone suggests that different risk measures may not only provide different point predictions but also a different ordering of individuals in terms of their risk preferences.

A second problem is best described using our nomenclature: First, for each of the risk elicitations, which are quite different from each other, one would have to use different experimental procedures. It may well be that the measurement of risk is not "just" a function of a participant's risk preference, but also of the *background noise **B***, perhaps mostly via *procedural noise* stemming from the experimental procedures used to elicit the risk preferences. This opens a whole new can of worms: Can we ever measure a person's "true" risk parameter?[57] Second, and somewhat related, it could be that risk preferences depend on the specific game or environment via the *background noise*, either via the *procedural noise* or via *other forces* or *models*, perhaps so far unknown forces. Puri (2024) provides a concrete example of one such *other model*, namely complexity, that affects risk preferences. She finds that "[p]articipants' risk premia increase as complexity increases, holding moments fixed […]". Puri (2024) hence suggests that levels of risk can be affected when we change the environment, specifically its complexity.[58]

Similarly consider trying to assess whether participants overestimate the expected value of the item conditional on having the highest signal, and whether this potential failure can help explain the overbidding in first price common value auctions. That is, we are looking for an estimate of the participant's beliefs about the expected value of the item we can "plug in" into our model to address its role in generating overbidding. I will describe in great detail three quite distinct ways to elicit computations of expected valuations. Those three ways *each* have their advantages and disadvantages. There is no "correct" answer on which one is best. That is partly what makes good design difficult. It depends on the conclusions you want to draw. While a specific point may have a correct measure, knowing the various ways may help you be aware of the assumptions you make to draw your conclusion.

One option is a measure most detached from bidding in auctions. We could measure how individuals compute valuations using a different signal structure, for a "*general abstract measure*" of a "failure to adjust." I hope that by now it is clear that using the estimate from this option may be problematic as it relies on the assumption that the failure to adjust is a trait we can measure in the absence of any *background noise*, or that the *background noise* in this measure is identical to the *background noise* participants experience while bidding.

A second option is to keep at least the parameters of the environment constant, that is, use a signal structure that is just like the one participants encounter bidding in the auction. We can ask participants, once they finished with the bidding rounds, in special estimation rounds, to concentrate on estimating the expected

[57] When trying to measure a person's "true" risk parameter, note that the implicit assumption is that a person has such a fixed underlying risk preference that informs all their decisions. Furthermore, for such a risk preference to impact all decisions in the same way, it is necessary that the participant has perfect awareness and perception about said risk, or, at least, for interpersonal comparisons, all individuals share potential distortions in the awareness and perception of the riskiness of the decision. This is the case not only for the situation at hand, but also for the risk-elicitation task. For example, an "old" unincentivized risk measure of psychologists had as one of the questions whether the person engages in the risky activity of skiing (presumably a "risky activity"). Being from Austria, where skiing is the national sport which almost everyone engages in from pre-school children to individuals who could be their grandparents, risky behavior and skiing do not have to go hand in hand.

[58] Note that distorted risk preferences may still be useful if at least the relative order is preserved. We have already discussed that this may not be the case across gender, but even beyond gender, Friedman et al (2013) finds that general measures of risk preferences often are not very correlated, see also Einav et al (2012). Furthermore, Puri (2024) finds that the role of complexity aversion on risk preferences is heterogeneous in the participant's cognitive ability.

value of the item for a given signal and signal rank. This second option, a "*related measure in isolation,*" keeps the environment from the auction and as such plausibly keeps the *procedural noise,* the part of the *background noise* resulting from describing and presenting the signal environment, constant. However, we may have changed the impact of other forces. For example, a participant may have more cognitive capacity when computing these valuations in an isolated decision rather than while bidding, and the latter hence may lead to larger computational problems and hence less adjustment to the correct valuation.[59] That is, the total *background noise* in these estimation rounds may well be quite different from the one participants experience while bidding.

The third option keeps the procedures from the auctions, and, in addition, doesn't remove other forces that may affect the computation while bidding: We can (in an incentivized way) ask for this valuation *while* participants decide what to bid. While not adding obvious changes to the value estimation (in case participants already engage in it), this "*related measure while bidding*" may add additional forces on *bidding*: For example, participants may now focus on having the highest signal. To avoid such a focusing on the highest signal, Lea Nagel, Emanuel Vespa and I have participants compute the expected value of the item conditional on having the median and the lowest signal rank as well. Additionally, a participant who just computed the expected value of the item for several signal ranks may have a "Eureka" moment which affects their bidding strategy by potentially reducing the winners' curse.

To summarize, let us consider general potential forces present in these three different elicitation procedures beyond those already present in the auction. Recall the *background noise* is comprised of *procedural noise*, to which we attribute forces stemming from the way a game is run and incentivized, and *other forces*, which are forces at work even if we were to use different procedures. For comparisons, we use as a default *background noise* the one participants experience when (potentially) computing the expected value of the item while bidding, since this is what we want to estimate (though, we actually estimate the expected value of the item conditional on having the highest signal). Note that *all three measure*s have the new procedure of describing the incentives used to elicit the valuation.[60]

1. "**General Abstract Measure**:" New *Background Noise*
   - *New procedural noise*: Need to describe the new game or environment with the relevant signal structure
   - *New other forces or models*: Different signal environments may also differ in complexity, potentially exhibit a different salience of numbers used in the task…
2. "**Related Measure in Isolation**:" New forces and missing forces
   - *Same procedural noise*: The description of the signal structure is like in the auction.
   - *New other forces or models*: Participants may be more exhausted if computing valuations at the end of the experiment and as such have less cognitive capacity which may reduce the ability to compute these valuations

[59] Alternatively, it could be that participants who place a high bid because they "enjoy winning" may justify this bid by an increase in the overestimate of the expected value of the item conditional on having the highest signal if this valuation is asked while bidding rather than in an isolated decision problem. It could also be that participants are in a general "bidding fever" which makes them overestimate the value of the items. Ngangoue and Schotter (2023) provide some evidence of a potential wedge between computations done in isolation or while bidding.

[60] A fourth option is to have participants compute valuations before bidding. One issue is that participants may not be yet as familiar with the signal structure, introducing a different noise or bias. Note that the related measure in isolation could also be done in between bidding rounds, the issues would be similar to when done at the end, however.

- *Missing other forces or models*: Potentially limited cognitive capacity to compute valuations while bidding (the opposite of the above), bidding fever…

3. "**Related Measure while Bidding**:" Same *Background Noise*, new force on *bidding*
    - *Same procedural noise*: The description of the signal structure is like in the auction.
    - *Same other forces or models*: Do not miss potential other models that apply to computing the expected value while bidding. Specifically, keep the potential force that bidders overestimate the value to justify their bid, that computations are done with limited capacity while bidding.
    - *New other models or forces on bidding!*: By forcing participants to explicitly compute the expected value of the item conditional on having the highest signal, they may adjust their bids downwards, perhaps leading to an Eureka moment.

When deciding which measure of expected value computation given the highest signal to elicit, it is important to keep in mind what we want to conclude, which question this measure is supposed to answer, why we want these value computations in the first place. Put differently, it is important to be aware of the identifying assumptions one has to make for the conclusions one wants to draw.

Note that if there is a presence of a new *force* on bidding due to the "Related Measure while Bidding," this suggests that participants are not computing such valuations on their own when bidding.[61] Which elicitation should Lea Nagel, Emanuel Vespa and I have used? If the question is how accurate, in principle, individuals are in computing those expected valuations, then the *general abstract measure* or the *related measure in isolation* might be most useful. If, in contrast, the goal is to assess how much a mistake in computing the expected value of the item – potentially amplified by cognitive constraints, bidding fever or any other force present in computations while bidding – affects bidding, then *related measure while bidding,* is the most adequate. Nagel, Niederle and Vespa (2025) are, as far as I know, the first to document that participants overestimate the expected value of the item conditioning on the highest signal. We use these computations to ask whether participants who bid more than the Bayesian expected value of the item conditional on the highest signal, also bid more than their potentially upward biased estimate of the expected value of the item. Put differently, do most participants bid *less* than their (biased) estimate of the expected value of the item conditional on having the highest signal? We find that the majority of participants do so. If we assume that participants bid using a (somewhat) symmetric Bayesian Nash equilibrium, that is, condition on having the highest signal, then it is reasonable to conclude that we found a, if not *the* major culprit for overbidding in first price common value auctions.

As a final "lesson" I hope this (maybe somewhat painfully) detailed description of the various measures makes it clear that no matter what measure we use, we make assumptions to justify their validity, and it is important to be aware of these assumptions. Furthermore, in general, there is no correct answer that applies to all situations: Like when running regressions, you need to know what it is you're controlling for, what the goal of the control is, and what assumptions you implicitly make when using said measure as a control. Finally, all of those measure constitute what I call *indirect controls*: We aim to estimate a force, like risk preferences, or biased valuation computations, which we then can use as controls in our model. Note that while I was very explicit on the assumptions made in order to take the results of each measure as pertaining

[61] To check whether bids are affected by valuation computations, Nagel, Niederle and Vespa (2025) compare bids when bidders are asked to compute valuations to those when they were not asked to do so. We find that bids are basically not affected.

to the problem at hand, we also make assumptions when we use the estimates from these measures in our econometric model, assumptions that often remain hidden. Basically, the assumption is that the model is well specified. The next Section will show that this assumption may be less innocuous than we might perhaps imagine.

3.4 Direct Control

In this chapter I discuss how to design your experiment when there are multiple potential models or forces you want to address, just like in the previous Section. I first discuss the general principle of a direct control. In Section 3.4.2 I provide an example where the implicit assumption inherent in an indirect control is violated, something we can demonstrate using a direct control. Section 3.4.3 provides examples of direct controls where indirect controls are especially problematic, namely when controlling for beliefs. This is because, in general, beliefs are complicated objects, and as such not easy to measure (though for a neat trick popularized by Mobius et al 2022, see Section 4.6).

3.4.1 Direct Control: Eliminate a specific force

An indirect control for a force (like risk aversion, or problems in computing the expected value of an item conditional on the highest signal) consists of trying to measure it and then using this measure to estimate its impact on behavior given a specific model that describes behavior. In the previous section I addressed some of the assumptions implicitly made when aiming to measure this force, and how those assumptions are minimized when the environment of the measurement mirrors the environment of the actions we aim to understand.

However, there is a second class of assumptions made when using a measure of the force when estimating its role. For example, in the previous Section, we assumed that participants, while bidding, condition fully on the highest signal. This implies that the relevant valuation that affects bidding is the expected valuation of the item conditioning on winning, which, in a symmetric equilibrium is the expected valuation conditioning on the highest signal.[62] Using this assumption, a large part of the winners' curse is driven by those biased valuations.

However, what if unusual mental models play a role? Suppose for the sake of the argument that participants, when bidding, condition on the signal being of median rank rather than being the highest of all signals. We don't know what signal rank participants condition on, theory only tells us what signal rank they *should* condition on in a symmetric Bayes Nash equilibrium. Suppose further that when estimating the role of biased valuations in accounting for the winner's curse, we use the bid function resulting from the symmetric Bayes Nash equilibrium. In that case, since we assumed that bidders condition on the median signal, the model is mis-specified! Even worse, our analysis will not help us discover whether the model is mis-specified. What to do? At this point it is important to remember that we are experimenters, we have tools at hand beyond econometrics: We can change the environment to directly address the role of specific forces.

Recall, in *design by elimination*, to show the importance of a new model m^* to account for behavior in a game g from Game $\boldsymbol{G}$ we aim to find a new Game $\boldsymbol{G}^{not\ m^*}$ and game $g^{not\ m^*}$ with similar *background noise*

[62] In fact, Nagel, Niederle and Vespa (2025) have a slightly different design, but for the sake of the argument, it is fine to think of it this way.

than game *g*, that has all the forces present in the main Game ***G*** apart from m^*. If behavior in game $g^{not\ m*}$ is different, this shows the importance of m^* in generating the behavior we found in the main game *g*. To design a direct control for some force *m*, we do something similar: We aim to find a Game ***G(not m)*** that is slightly adjusted from ***G***, but in which *m* does not affect behavior. If behavior in the game *g(not m)* mirrors that of *g*, and the two games have similar *background noise*, then *m* was not integral for the behavior in *g*. If behavior is different, then we can use that change in behavior as a measure of the role of *m*.

3.4.2Example of a Direct Control

What does a direct control for computational problems in computing the expected value of the item conditional on various signal ranks look like? In Nagel, Niederle and Vespa (2025) we provide participants, while they place a bid, with a table that, for the given signal the participant received, computes the expected value of the item for that signal for each signal rank. Hence, in these rounds, we eliminated, *by design*, the role of computational problems when computing the expected value of the item. Given the conclusions we drew from the indirect control exercise, we expect participants to significantly lower their bids when they receive this information, that is, be less prone to fall prey to the winners' curse. However, we find that the bids of participants, whether they bid with or without this table are virtually identical.[63] This direct control therefore delivers a *completely different conclusion* than the one we drew from the indirect control!

What can account for this difference in results from an indirect and a direct control? In Nagel, Niederle and Vespa (2025) we show that the discrepancy arises from the fact that the implicit assumption used to econometrically control for those computational errors is not justified. Specifically, when participants place a bid, they *do not fully* condition on having the highest signal. Put differently, it is true that participants overestimate the expected value of the item conditional on having the highest signal. Furthermore, any such bias, if used to explain bidding in a symmetric Bayes Nash equilibrium where bidders condition on the highest signal immediately accounts for a fraction of the winners' curse. However, the assumption that bidders condition on the highest signal is not justified, and as such the econometric model that allows us to use the biased estimate conditional on the highest signal is mis-specified!

For the analysis that allows us to show that most participants do not condition on having the highest signal when placing a bid, I invite the interested reader to consult Nagel, Niederle and Vespa (2025), as there are multiple design components I have not mentioned here that are important for such a conclusion. In our paper we also address which combination of models is most relevant when accounting for the winners' curse. Beyond being necessary to draw the right conclusions, understanding what affects bidding and hence delivers the winners' curse is important if we want to understand why individuals struggle in this environment, and what we could potentially do to help them. It also helps us understand which of the many potential behavioral forces we have identified over the years as behavioral economists that could apply in this environment are relevant and do apply. Finally, we show that a single force cannot explain everything, rather it is a combination of forces or models that matter.

3.4.3 Direct Control: Beliefs

[63] Note that one possible reason for a lack of change in bids could be that participants do not pay attention to the table (maybe it is in the wrong format, too complicates, not easily visible on the screen), in which case the table only provides a control for computational problems in theory but not in practice! We, however, show that this is not the case, see Nagel, Niederle and Vespa (2025) for details.

I provide two more examples of direct controls in an area where they are perhaps especially useful because indirect controls are especially complicated when we aim to control for beliefs. The first example is from Nagel, Niederle and Vespa (2025) and concerns beliefs players have about others, and the second example is from Exley and Kessler (2022) and concerns beliefs about the participant's own performance.

Unusual mental model explanations for the winners' curse include, beyond a failure of contingent thinking, cursed equilibrium (Eyster and Rabin, 2005) and level-k thinking (Crawford and Iriberri, 2007). An interpretation of both models is that participants have non-equilibrium beliefs about the behavior of others and best respond (in Eyster and Rabin, 2005, to some extent) to these unusual or non-equilibrium beliefs.

One can view these models as a description of the data, or take them literally, that individuals do hold such unusual beliefs. Clearly, asking participants to provide beliefs about the behavior of others in an incentivized way is complicated, as beliefs are complicated objects. A tractable alternative is to fix the behavior of others, inform participants about this fixed behavior, and then test whether participants best respond to their knowledge about others' strategies. What should such a behavior of others look like?

One option is to explain, for example, the symmetric Bayes Nash equilibrium strategy to participants, or any other strategy, and then check whether participants best respond to this strategy. However, such a design has the clear drawback that bidders are told about a bid function they may or may not have thought of themselves. Providing such a "new" strategy, on its own, can either cause a Eureka moment, or participants may simply mimic the provided strategy, infer more from its provision than perhaps is warranted.

In Nagel, Niederle and Vespa (2025) we use a second approach – one I have already used in Ivanov, Levin and Niederle (2010) – which does not suffer from having to provide participants with a strategy they have not thought of themselves: We use the method I describe in more detail in Section 4.4. We first elicit the bid function of each participant. We then inform participants that, instead of bidding against 6 other players, they bid against six computers who use the participant's own past elicited bid function (and we even remind them what that is, to really fix the beliefs of the participant about the behavior of others). This implies that bidders (and the experimenter) have full knowledge of the bidders' beliefs about the strategies of their opponents without, however, teaching bidders anything about possible bid functions they may not have thought of themselves! In Nagel, Niederle and Vespa (2025), as in Ivanov, Levin and Niederle (2010), we find that, by and large, participants do not change their bids whether they bid against others or their own past bid function. This is the case despite the fact that most bidders do not use a strategy that is a best response to their own past bid function.[64] This suggests that the explanation that the winners' curse is the result of participants best responding to unusual beliefs about others' strategies is likely not borne out in the data.

A second example of a control for beliefs about one's own performance, and why it is important to directly control for it, comes from Exley and Kessler (2022). They have participants take a math and science test with 20 questions, and then answer a series of subjective self-evaluations about that test, like whether they agree with having performed well on the test on a scale from 0 to 100. They find that women evaluate

[64] There are several reasons why this might be the case: First, bidders may not know how to best respond, which seems to have been the case in Ivanov, Levin and Niederle (2010). In Nagel, Niederle and Vespa (2025) bidders do react with their bid function to some information, just not when it comes to information about the strategies of others. This could be either because it is too complicated, or because participants already believed others behave like them, but then, largely, fail to best respond to those beliefs.

themselves less favorably than equally performing men do. One possible explanation is a gender gap in beliefs (or confidence) about one's own performance.

To assess the participant's beliefs about their performance, they ask them in an unincentivized way "out of the 20 questions on the test […], how many questions do you think you answered correctly?" If payments were a fixed sum if they are (close to) the truth and zero otherwise, it would incentivize providing the modal of one's distribution over possible number of correct answers. With this indirect control, using this – arguably not complete – measure of beliefs about one's performance, they find that the majority of the gender gap in self-evaluations appears to be "accounted for" by these beliefs, suggesting that a gender gap in self-promotion is basically a gender gap in beliefs about one's performance.

However, Exley and Kessler (2022) also use a direct control: They inform participants about both their absolute and their relative performance before eliciting self-evaluations. They still find a large and significant, albeit somewhat reduced, gender gap in self-promotion. This gender gap, by design, cannot be attributed to gender differences in beliefs about one own absolute or relative performance!

As in Nagel, Niederle and Vespa (2025) the direct control provides a different conclusion from the indirect control. Christine Exley and Judd Kessler discuss, albeit only in the appendix, that controlling for beliefs statistically may result in measurement error, reverse causality, or omitted variable bias that could arise, e.g., from the existence of types of individuals who broadly view their performance more positively across various types of questions and as such also have a higher belief in their performance. The strength of controlling for beliefs directly, that is by design, is the ability to sidestep such econometric complications, such as an omitted variable bias or a mis-specified model.

3.5 New Comparative Static or Do it Both Ways

The design features I presented so far involve, in some way, an element of elimination. In a *design by elimination*, the idea is to construct a very similar environment to the original one, where, however, your new model m^* is expected not to have an impact. You therefore construct a counterfactual without m^* to demonstrate the role of m^* in the original environment. In a *direct control*, you eliminate the "nuisance force" in question and demonstrate the impact it has on your original result. In both cases we essentially create alternative universes that are very very similar to the original one, but in which a specific force has no power to affect decisions of participants. We then use this alternative universe to show the role (or lack thereof) of this specific force in generating the original result.

However, sometimes it may not be possible to fully eliminate a force. We can then, instead, aim to mitigate it, or create alternate universes where the impact of the force we want to focus on differs from all other forces. Such a design is a "*new comparative static*" or "*do it both ways*" design. Specifically, in an extreme way, we may be able to generate a new environment or game that is very similar to the original one, but where the force we want to control for pushes actions in a different way than all other forces. This is basically the idea behind the "old school comparative static" design as well, see e.g. Kagel and Levin (1993). However, the difference is really in ensuring that the new environment or game is very similar to the previous one, such that we can confidently assert that the *background noise* should affect participants similarly in both games. This is especially relevant if the alternative hypothesis is not "just" not-neoclassical. I present this logic in Section 3.5.1 and provide a very ancient example of such an experiment in Section 3.5.2. For a more recent example, see Section 3.5.3.

3.5.1New Comparative Static Design

Consider the case where a *design by elimination* to provide evidence that the model m^* is responsible for the outcome in game *g* is not feasible, or we simply look for an additional way to test for the role of m^*. We can use a new ***comparative static*** design: Find a game *g'* where (i) model m^* differs in their prediction of behavior from the alternative model, (ii) the expected behavior of m^* in game *g'* is quite different from that of m^* in game *g*, and (iii) the expected effect of the *background noise* in games *g* and *g'* is similar.

For point three to be plausible, game *g'* needs to be sufficiently similar to game *g*, something that was not always the case in old comparative statics experiments. Point two rules out the following scenario. Suppose the prediction of m^* in game *g'* is similar to the prediction in game *g*, while the alternative model predicts a change in behavior. If the behavior in *g'* follows the prediction of m^*, this could be entirely dues to a strong effect of the *background noise* on behavior in both games rather than an effect of m^*. Considering point one, one version of such a comparative static design is where the alternative model makes the same prediction in games *g* and *g'*. Alternatively, it is actually sufficient if the alternative model makes a different prediction in game *g'* than m^* does. This may, however, cause problems, if the qualitative changes between *g* and *g'* are identical for the alternative model and m^* and they only differ in their quantitative prediction. This is because the background noise may complicate precise quantitative predictions.

3.5.2An Old Example of a New Comparative Static Design

Probably one of the earliest well-documented experiments which uses a “Do it Both Ways” design is one I learned from Alvin Roth in his graduate Experimental Economics class at Harvard: While the result of an initial experiment may be due to ones' favorite hypothesis, it may be due to other reasons one has not thought about (*background noise*). To test the favorite hypothesis, change a minute detail that *only* affects the behavior if the favorite theory is correct, but likely not if the result is due to any other possible hypothesis. The results of this new treatment then either confirms or disproves the preferred hypothesis driving the experiment in the initial game.

This early example is from the book of Judges of the Old Testament, Chapter 6. The story begins with the Israelites, turned away from God after 40 years of peace brought by Deborah's victory over Canaan, being attacked by the neighboring Midianites. God chose Gideon, a young man from an otherwise unremarkable clan from the tribe of Manasseh, to free the people of Israel and condemn their worship of idols. Gideon, to convince himself that the voice he hears is indeed the voice of God, rather than hearing voices for other reasons, asks for a test:

> And Gideon said to God: 'If You will save Israel by my hand, as You have said look, I will put a fleece of wool on the threshing-floor; if there be dew on the fleece only, and it be dry upon all the ground, then shall I know that You will save Israel by my hand, as You have said.' And it was so; for he rose up early on the next day, and pressed the fleece together, and wrung dew out of the fleece, a bowlful of water.

So far, this makes Gideon a decent experimenter (testing whether voices are due to God), but not yet a great one. Gideon, indeed, realized, that there could be alternative explanations for the main result. It could be that this is simply the way it is, after all, he probably wasn't used to leaving a fleece outside. He might worry that his design does not distinguish between whether the outcome is due to the God almighty, or merely the way cold nights interact with a fleece left outside. You can see that as a believer, a *design by elimination* would be difficult to conceive. Hence, Gideon opted for a comparative static design. In his next design all other hypotheses would generate the same result, as the *background noise* is held constant, but, if the initial result was due to God, we would expect a change in outcomes. Here is his minute design change that does not affect *background noise*: He asks God whether he can reverse the result:

> And Gideon said to God: 'Do not be angry with me, and I will speak just this once: let me try just once more, I ask You, with the fleece; let it now be dry only upon the fleece, and upon all the ground let there be dew.' And God did so that night; for it was dry upon the fleece only, and there was dew on all the ground.

For all empirically minded readers, it turns out that this experiment, at least according to the old testament, has external relevance: This is because Gideon raised the army which indeed defeated the Midianites, as was promised.

3.5.3 A New Example of a New Comparative Static Design

For a more recent example, we return to Martinez-Marquina, Niederle and Vespa (2019), where we showed that when a participant submits a price to one firm that is either of low value 20 or high value 120, each equally likely, then such a participant is likely to submit a price of 120 and hence make profits equal to a 50:50 lottery of -90 (if the firm is of value 20) or 60 (if the firm is of value 120) rather than a price of 20 and hence receive a 50:50 lottery that pays either 10 or 0, respectively. To focus on the role of computational complexity only and eliminate risk, we invented the deterministic treatment, where the participant submits a price that applies to both firms, the 20 and the 120 firm, which each exist. We find that participants are then much more likely to submit a price of 20 for a certain profit of 10 than a price of 120 for a certain profit of -30.[65]

Our favored interpretation is that the deterministic treatment measures the computational complexity of the problem, since in both treatments (the probabilistic and the deterministic treatment), participants have to consider both firms to compute the payoff maximizing price. The power of certainty, the increase in expected payoff maximizing choices between the probabilistic and the deterministic treatment, shows how much harder decisions are when states are hypothetical rather than certain. We speculate that participants, when firms are hypothetical (or potential) rather than realized firms, have a hard time to focus on both firms at once, but they have no such problems when the firms both exist.

However, one can always come up with other ad hoc reasons why participants may be more likely to submit a price of 20 in the deterministic treatment that is not due to an increase in rational, payoff-maximizing choices. For example, maybe participants want to only buy one firm (to make at least some money) and don't want to think of a second firm (for example because thinking is costly). This is a possible explanation why, in the deterministic treatment, they submit a price of 20, which happens to be also the payoff-maximizing action.

We therefore introduce a small variation in the design to provide additional evidence for our interpretation. This helps confirm that choices in the deterministic treatment are indeed more likely to be payoff

[65] To keep expected payments constant across the two treatments, each currency unit in the deterministic treatment was worth half of the currency unit from the probabilistic treatment.

maximizing rather than driven by a desire to only buy one firm. We introduce rounds (in both the deterministic and the original probabilistic treatment) where the values of the two firms are closer together, such that 1.5 times the value of the low firm is actually strictly larger than the value of the high firm. This implies that now submitting a price equal to the high firm is the dominant strategy. Hence, if results were due to our favorite hypothesis (*m^**) that participants are (more) payoff maximizing, we expect a change in behavior. In contrast, forces that consist of having participants only buy one firm, do not. A participant who submits a price of 20 when the firm values are 20 and 120 for the "wrong" reasons, like a desire to only buy one firm, is not expected to switch the price to the high firm in these new rounds. To be cautious when classifying participants by whether they are payoff maximizers, we only consider participants who submit payoff maximizing prices when the two values of the two firms are far and when they are close to each other.

3.6 Stress-Testing and "Do It Both Ways" Experiments

I think from all the previous example one message is clear: It is, unfortunately, possible to receive results for the "wrong" reasons, which then affects the validity of the conclusions drawn from experiments. In the end, there is no perfect way around this, we can only do our best to control for other forces as much as possible. Furthermore, across the board, we assumed that the *background noise* does not interact with our preferred hypothesis m^* in a *design by elimination* or a *comparative static design*, and other models when many models may drive results. Sometimes this assumption may be somewhat heroic. For all these reasons, we sometimes may want to include a ***stress-test*** of the main hypothesis. If we believe that we found convincing evidence for a specific model, or, more precisely, convincingly ruled out alternative explanations for a given phenomenon, then we can still try to include a stress-test. A stress-test consists either of checking auxiliary predictions given the data at hand, or, even better, checking predictions in a slightly new environment where, if the new model is correct, we know what the predictions would be, predictions that ideally differ from other models.

While at face value this looks like a comparative static experiment, the difference is that here we add those stress-tests as a "final check" of our hypothesis. Furthermore, stress-tests can consist of completely different environments created specifically to provide additional evidence for the main hypothesis. In the examples throughout this survey, I have already talked about stress-tests people have been doing. I will provide you with one final example from my paper with Alejandro Martinez-Marquina and Emanuel Vespa.

In Martinez-Marquina, Niederle and Vespa (2019) we provided a lot of evidence that participants are much better at finding the payoff maximizing action when there is no uncertainty involved. Our intuition is that participants are more likely to actively think of both possible states of a firm, a firm of value 20 and one of value 120, when the two states or firms exist and are realized states, rather than when they are only hypothetical or potential states.

The stress-test for this intuition involves a new environment where we ask participants to give advice to another participant on what price to submit. We assess whether participants mention all four possible outcomes, that is submitting a price of 20 when the firm is of value 20 or 120, and the two outcomes from a price of 120, respectively. We find that participants are economically and statistically much more likely to mention all four outcomes when both firms exist. That is, it seems as if they are indeed more likely to think of all states when states are certain rather than hypothetical. Furthermore, adding a dummy on whether all four outcomes (two states and two prices) are mentioned reduces the differences in the propensity to submit expected profit maximizing prices between the two treatments.

4. Advanced Design Tools

In this last subsection I describe a handful of advanced design tricks and techniques that are at times deeply rooted in theory and that you may not have come across in a standard Experimental Economics class. Note that I am not advocating that you have to use them in all instances. However, when you make some specific theoretical claim, it better be the case that your design backs the theory rather than renders it invalid! Overall, I think it is useful to have a few tricks up your sleeve in case you have some specific design needs.

4.1 Infinitely Repeated Games

While it is clear how to implement finitely repeated games or decision problems, there is a way to implement infinitely repeated games where payoffs in subsequent rounds are discounted by a factor $\delta<1$. To do this, after each round of play, instead of discounting the future payoffs by a factor $\delta<1$, you only implement the next round with a probability $\delta<1$. That is, in each round, if you reach that round, instead of the chance of a next round happening with certainty (or never), the next round happens with probability $\delta<1$. This random termination method was introduced by Roth and Murnighan (1978).[66]

If you want to compare the behavior of participants in such a game or maximization problem to the Bayes-Nash equilibrium or the payoff maximizing action, there is one more important point, namely, how to compensate participants for their behavior. Outside of the laboratory, or in the model, participants consume their payoffs in every period. However, in the laboratory, we, in general, pay them at the end of the experiment.[67] Roth and Murnighan (1978) have payments that depend on outcomes in all rounds (perhaps slightly more so even in Murnighan and Roth, 1983, as their first paper has quite some deception).

Chandrasekhar and Xandri (2023) show that paying all periods is valid only when agents are assumed to be risk neutral (though, as they point out, some researchers pay every round even if the model they want to test involves an explicit deviation from risk-neutrality).[68] While in many experiments researchers pay a random round and this is often the cleanest way to preserve incentives (see Azrieli et al., 2018), you should really not use this when deciding which rounds to play in an infinitely repeated game. The reason is that it puts relatively more weight on early rounds and as such distorts incentives. To preserve incentives, you should pay the final round (for details see Chandrasekhar and Xandri, 2023). This point was also made by Sherstyuk, Tarui and Saijo (2013) who provide evidence that paying a randomly selected round distorts behavior compared to paying all rounds or only the last round.

4.2 Inducing Risk-Neutrality

[66] Whenever you present a paper using this method, you are likely to have at least one theoretically minded economist who will share their worry that, possibly, one of those games may indeed last forever. Just so you know and can't say no one warned you.

[67] The reason I say in general is because there are some papers that involve consumption (see Augenblick, Niederle and Sprenger, 2015), though as far as I know not in infinitely repeated games yet.

[68] For additional thoughts on payment schemes see also Azrieli et al. (2018). They further document that of the papers they survey, almost half do not justify their payment scheme.

An early bargaining model still used in some parts of economics is due to Nash (1950), where two bargainers are faced with a set A of alternatives. If they both agree on some alternative α in A, then α will be the outcome. Otherwise (i.e. if they fail to agree on an outcome) the result is a fixed disagreement outcome δ. Let u_1 and u_2 be the expected utility functions representing the preferences of player 1 and player 2. Let S be the set of feasible utility payoffs from an agreement, i.e. S = {x = ($u_1(\alpha)$, $u_2(\alpha)$) | α is in A} and let d be the utility payoffs to the players from a disagreement, i.e. d = (d1,d2) = ($u_1(\delta)$, $u_2(\delta)$). Under the complete information hypothesis, that is when (S,d) is known to both bargainers, the Nash bargaining solution (which is (i) invariant to affine transformations, (ii) pareto optimal, (iii) independent of irrelevant alternatives and (iv) symmetric) is the set of points x=(x1, x2) that maximize (x1-d1)(x2-d2).

But how can we make (S,d) known to both participants? How do we, as experimenters, know how they translate payments (if we use monetary payments in the experiment) into utils? Early experiments assumed linear utility in money, that is $u_1(\alpha) = u_2(\alpha) = \alpha$. Consider participants who divide a fixed amount k. Now consider, in addition, that a dollar to player 1 "costs" a from the fixed amount k, and a dollar for player 2 costs b, that is A = {(x1, x2)| ax1 + bx2 ≤ k}. If they cannot agree, both players receive 0. Early experiments were interpreted as if (S,d) = (A,0). Such experiments found a lot of equal divisions, even when a and b were different.

However, this kind of design turned out to be unpersuasive to game theorists, who felt that an experiment that *assumes* everyone's utility function is the same cannot do justice to a theory that models *all* individual differences as differences in utilities. So, Roth and Malouf (1979) introduced binary lottery games. Basically, consider a large (L) and a small (s) prize for each player, the set of alternatives is then A = {All lotteries over ($L1, $s2), ($s1, L2)} and the disagreement point gives both players the small prize: (s1,s2).

Note that if v is an arbitrary utility function for money, and q is a lottery that gives the player a probability q of receiving $L and (1-q) of receiving $s, then we can use an affine transformation to represent this utility as u(q) = q (where v(q) = aq + b, with b=v(s) and a=v(L)).

Now, (S,d) is (up to affine transformations of each axis) the convex hull of (0,0), (1,0), and (0,1). Furthermore, assuming rationality, (S,d) is known to both participants! Note that the binary lottery game design controls for the predictions of the theory, but is not necessarily a control for the behavior of the bargainers…

The complete information hypothesis (and the auxiliary assumption that players are utility maximizers with utility functions that have as arguments only their own payoff) now predicts that information about the prizes won't affect the bargaining outcome! Roth and Malouf (1979) show that this is not the case, confirming "fairness" motives of bargainers. While this was a controversial result at the time, since then, of course, fairness as a concept in Experimental Economics, and in Economics in general, has become, I am going to say, non-controversial.

The main reason for this little history lesson is that using binary lotteries is a way to transform an individual's utility function into a linear function. Note that this, obviously, assumes no probability weighting! It is perhaps a slightly flawed methodology, but in case you speak to neoclassical economists who worry about curvature in utility, this is a way to address this concern.

4.3 Eliciting Beliefs

Beliefs, which are often instrumental to selecting an action, are notoriously hard to elicit. This is because beliefs about an outcome can be a complex function, and eliciting functions is hard. In Mobius (2022) we (re)discovered and used in experiments we ran early in this millennium that one way to make beliefs tractable is to consider binary (or finite) categories.[69] For example, consider an individual who performs in a short IQ-style test. One performance question might be "How many problems do you think you solved correctly?" If there were twenty questions, this may require eliciting 20 numbers (which add up to 1, hence we can infer the 21st), namely the chance I solved 0 problems correctly, 1 problem correctly, etc.

As this is cumbersome, some researchers instead ask participants to report how many problems they think they solved correctly and promise some positive payments if they are right. In this case, the participant is incentivized to provide the modal of her belief distribution, that is, the number of answers she is most likely to have solved correctly.

If we do need to have precise beliefs, perhaps to study Bayesian updating, we can turn the problem into one of fewer categories, by, for example, asking: "What is the chance that your performance was in the top 50% of participants."[70] Beliefs are now one number that we can elicit. In Mobius (2022) we used what we called a cross-over mechanism: Varying the given probability, we ask participants whether they prefer to be paid a fixed prize with some given probability, or whether they want to be paid the prize when their performance is in the top 50% of participants. By varying the chance to receive the fixed prize, we elicit at what chance participants believe receiving the fixed prize at that chance is as good as receiving the fixed prize when they are in the top 50% of performers.

This elicitation method does not rely on individuals having linear probabilities (that is, can accommodate probability weighting) and also has another advantage, often overlooked in the literature: Suppose we ask participants before their performance about their chance to perform among the top 50% of performers (as we do in our experiment). Let's say the participant said they have a 70% chance to be in the top 50%. Suppose the participant then performs the task, does the participant who performs better than expected face a trade-off between more monetary payments for the performance, but lower monetary payments for overshooting their stated beliefs? When using the cross-over mechanism there is no such trade-off, it remains in the participant's best interest to keep solving problems correctly, for details see Mobius (2022).[71]

4.4 Turning a Game into a Decision problem (controlling for Beliefs)

In many games decisions depend on the participant's beliefs about the behavior of others, which is why as market designers we like implementing outcomes via dominant strategies, whenever possible. Suppose we observe actions of an agent in a game without a dominant strategy, that is where their beliefs might affect their choices. Suppose the action does not correspond to one expected under the neoclassical model. Is the deviation due to unusual preferences, frictions or unusual mental models, to use the nomenclature from

[69] Our mechanism is equivalent to Grether's (1992) BDM probability pricing procedure. It has also been independently proposed by Allen(1987), Karni (2009) and Holt (1986).

[70] It may be important, in case it is not obvious from your experiment, to point out who exactly the other participants are whom the person should compare themselves to.

[71] Suppose the participant performed to have a 90% chance to be in the top 50%. Suppose the participant is paid for their beliefs and the randomly drawn chance is 60% for the lottery prize: Then the participant stated they prefer to be paid if they are in the top 50%. Instead of receiving the prize with a (stated) 70% chance the participant receives it now with 90% chance, a win for the participant. If the randomly drawn chance for the lottery prize is above 70%, the participant gets paid based on the lottery and their performance doesn't affect winnings.

Nagel, Niederle and Vespa (2025)? One unusual mental model is that the participant has non-equilibrium, or unusual beliefs about the strategies of others. Such unusual, or non-equilibrium beliefs can cause havoc with the neoclassical model. Recall, for example, the famous gang of four paper Kreps et al (1982), or models of *level-k* thinking (Nagel 1995, Costa-Gomes and Crawford 2006) or cursed equilibrium (Eyster and Rabin, 2005, and Cohen and Li, 2024).

How can we infer the role of those beliefs? Unfortunately, when a player's action depends on their beliefs about the decisions or strategies of others, it is difficult to elicit those beliefs when the action space is large, see the previous subsection. In Ivanov, Levin and Niederle (2010) we use a technique that simultaneously eliminates the role of beliefs, without, however, teaching the participant anything about possible strategies the player may not have thought of herself: We turn the game into a maximization problem. For symmetric games, like the two-player auction in Ivanov, Levin and Niederle (2010), we first elicit an (incentivized) strategy of the participant. We then inform the participant that their opponents have been replaced by computers who all use the participant's own elicited strategy.

In such an environment the experimenter has full control over the participant's beliefs about the strategies of their opponents, you can even remind players what that strategy is, just to be sure. We use this strategy also in Nagel, Niederle and Vespa (2025) where we conclude, like in Ivanov, Levin and Niederle (2010), that unusual beliefs do not play a large role in accounting for behavior that deviates from the symmetric risk-neutral Bayes Nash prediction. In simple two-player guessing games, Fragiadakis, Knoepfle, Niederle (2024) use this method for asymmetric games, and show that a sizable fraction of participants does change their action to best respond to their own past behavior (even when not reminded what said behavior was). That is, in two-player guessing games, but not in arguably much more complicated auctions, unusual beliefs about the behavior of others may account for the behavior of participants.

Basically, turning the game into a maximization problem allows us to eliminate the role of unusual beliefs, and helps us ensure that we know the participant's beliefs about the actions of their opponents. It allows us to compute the best response, without having to assume that the participant holds specific beliefs, which may or may not be a justified assumption.

4.5 Comparing Apples to Apples (and not to Oranges)

In some experiments we want to examine whether the impact of changes in payments in context 1 mirror the impact of changes on payments in context 2. One important aspect to keep in mind when making such comparisons, is that the stakes in context 1 and context 2 are comparable. Let me give you a concrete example. Exley (2016) examines how much individuals dislike risk when the money is for themselves, compared to when the money is for a charity. Specifically, she has four comparisons. In two of them, participants are fairly risk-neutral, namely when participants make decisions between lotteries and certain amounts that are either both in the self (own money) domain, or in the charity (money for charity) domain. However, when one payment is in the self domain and one in the charity domain, then behavior become interesting. For example, when evaluating a lottery for charity, participants appear extremely risk averse and are willing to give up this charity lottery for relatively little certain money for themselves. This behavior mimics a participant who uses risk-preferences as an excuse to behave selfishly.

When choosing the payoff amounts in these risk lotteries, Exley (2016) could have set self lotteries to yield \$10 for participants with some chance and charity lotteries to yield \$10 to some charity with some chance. Doing so might already have resulted in differences in responses to risk even without any trade-off between own and charity payoffs. This is because almost everyone values $10 for themselves more than $10 for charity. Hence, by having the payoffs as $10 for oneself and for charity, we would introduce, by design, a difference in stakes between decisions that involve only one's own payoff and decisions that involve only charity payoffs. Such changes in stakes could, in turn, have led to wrong conclusions about the extent of evidence for excuse-driven behavior.

In experiments, we must compare "apples to apples" to the best extent possible. All features can never be held perfectly constant across context. However, some features, like stakes(!), are known to matter quite a lot and should be taken into account. Exley (2016) adjusts for the role of stakes by using a multiple price list to estimate an X value such that participants are approximately indifferent between $10 for themselves and $X for charity. She then examines participants' responses to self lotteries involving $10 to participants with some chance and charity lotteries involving $X to charity with some chance. It is only by changing payoffs in the two domains that we can keep stakes comparable across the two environments.

Exley and Kessler (2024) employ a similar strategy but push the "apples to apples" comparison even further by focusing on decisions that are exactly the same across their self/self context and charity/self except for one payoff option and that one payoff option is calibrated such that participants value it approximately the same in both contexts.

4.6 Testing Expected Utility

Coming back to how to pay participants, and understanding what assumptions are made, I highly recommend looking at Azrieli et al (2018). They basically show that paying one randomly selected decision is incentive compatible under an often-mild monotonicity assumption. However, Segal (1990) in Theorem 2 shows that this monotonicity assumption (termed "compound independence") together with reduction of compound lotteries implies the familiar independence axiom of expected utility theory.

This suggests a word of caution: If in your model you assume reduction of compound lotteries and you also pay one random decision to provide proper incentives (and thereby assuming compound independence) then you implicitly assume that preferences satisfy independence. This implies that if you observe independence violations in your experiment but assume reduction of compound lotteries, then compound independence must fail and so may incentive compatibility of the experiment (see Example 3 of Azrieli et al., 2020 for discussion).

It is maybe not surprising that overall, I spent a lot of time on discussions on how to pay participants. Personally, I think it is very important to be aware of your assumptions, and knowing whether they are plausible, or whether you violate a specific model. I think a good general rule is to design your experiment in a way to avoid discussions that the decisions you made in theory affect decisions, even if you think that in practice they do not.

5. Conclusion

A final topic that warrants attention concerns how much we can trust the results of a paper and a whole literature. While related, those two trust issues differ in important ways, a difference that I worry is sometimes underappreciated. There has lately been a huge push to propose or require policies that are geared towards increasing how much we may trust the results of a paper. I will first define (very briefly) what these are. I then describe the potential problems of a paper and a whole literature and whether problems are (in expectation) smaller or exacerbated in experimental economics compared to other empirical methods. Finally, I discuss the possible effects of various policies on each of the potential problems a paper or a literature may suffer from. I will end this section with my views on the current state of affairs (though to be fair, a big chunk of this section already represents my views that are not always aligned with editors from famous journals).

A *pre-registration* of a paper is best thought of as describing the goal of the experiment, a broad description of the design, including perhaps the number of participants one aims to collect data from, the main hypothesis and how the design parts broadly address those hypotheses. A *pre-analysis plan* is in essence similar, but much, much more detailed. The idealized version is basically to almost to write the paper in advance, with the econometric specifications, tables and figures of your paper, so that, after collecting the data, the hands are basically completely tied as to what analysis is included in the paper. Note that, however, as far as I know, any proponent of pre-analysis plans acknowledges that, as an author, one should be allowed to do additional analysis, though this should then be flagged in the paper. A *replication* can take many forms, the least expansive is being able to reproduce the results from the paper with the given data. An *exact replication* is re-running a given experiment using the original instructions and a replication in a broader sense is almost akin to a robustness test, reproducing results while changing non-essential (given the model or claims from the paper) parts of the environment or game, instructions, etc.

To discuss the potential issues plaguing research on experimental economics, I will start with concerns on the paper level and then move to the literature level. To provide a clear idea of the problem, I discuss in each case what kind of replications (in potentially a very broad sense) are necessary to alleviate the specific concern. The potential problems one might have with a paper/literature are

1. *Given the game (or environment) and the data, can we trust the analysis?* There are two levels of concerns. One is that there were innocuous errors in the data analysis. The second refers to issues related to *p*-hacking and multiple hypothesis testing, where researchers describe only a subset of analyses, focusing on those that yield significant results and hiding the others.
2. *Given the game, can we trust the data collection?* This questions the method of collecting the data, which could range from continuously collecting data until receiving a significant outcome, and issues of data being "contaminated," perhaps by experimenter demand effects.
3. *Is the paper a test of the model or does the model "just" describe the data?* This questions the interpretation of the result, even when researchers did not use questionable research practices when collecting or analyzing data.
4. *Given the results of the paper, how robust is this conclusion?* Even if we trust that points 1-3 are fulfilled, what do we think about the robustness of the conclusion? If the paper is describing an anomaly, a specific new insight, we may understand that the authors potentially went through great lengths to find an environment or game to achieve the desired outcome. Alternatively, the paper may make a point that is described as a robust general result. To trust this robustness and perhaps even the validity of the result, we may worry how much the researcher engaged in *g*-hacking. Specifically, did the author engage in extensive undocumented pre-testing of the environment for reasons beyond those described in the paper? Even for a given game structure or environment, did the author engage in extensive pre-testing to check whether the specific parameters, the specific experimental instructions, the *benchmark noise* in general contributes to the result in a way that is not described in the paper?
5. *Given the published papers, can we trust the conclusion of the literature?* There are two reasons we may not trust a literature, beyond issue 4. One is publication bias, the fact that positive results

> are more likely to be published. Even if every research were to study a simple effect, given that there are myriad effects and many researchers, some may receive results that later will turn out to "have been lucky," even if the researcher in question does not engage in any shenanigans. Finally, if the literature did show that a given result is replicable, can we trust that the literature has replicated not only the very specific result, but also expanded the validity of results to environments where, given the claims of the papers, we expect the result to apply?

The issues 1-3 operate on a paper-level. Issue 4 and 5 are more on a literature level. Issue 4 is about the robustness of the result, which can be made to appear broader than may be warranted and hence be more questionable due to activities of the author. As such, it does not necessarily question the results of the paper, but rather whether they are as general as they are presented to be. Issue 5 concerns trust in a finding on a literature level.

Note that as a literature, or as a scientific collective, I think we should really care about all points, all the way to the fifth one. As a journal, it could be that the worry is more concentrated on the earlier points. For example, suppose the literature is very vibrant and quick to replicate "important" or well-published results that have a lot of citations, then such a literature may be relatively "less worried" about issues 1-3 but in turn be concerned that the scientific environment fosters engagement with issues 4 and 5. In a way, consider a paper published in a vibrant and active literature with strong incentives to investigate the robustness of important results. If the paper is important, it will be investigated which implies it will quickly be evident whether the authors engaged in questionable research practices or were surprisingly lucky with their data. A vibrant or fast-moving literature may hence find it more important to ensure continues incentives for issue 5 and 4 rather than jeopardize such incentives "just" to take care of issues 1 to 3. This is not because these issues are not important, it is just that even if they were take care of, we would need to ensure we take care of issues 4 and especially also 5, the latter of which is almost unavoidable for a single paper!

Hence, when deciding about policies, it will be important to assess not only how much they affect issues 1-3, but also whether they impact issues 4-5. If there is a trade-off, I personally rather err on the side of the literature rather than the side of the paper and the journal!

Before discussing the possible impact of various policies, I consider the evidence of problems in the experimental economics literature. While, strictly speaking, it is not relevant whether experimental economics suffers more or less than other applied fields from producing results that cannot be replicated or are not robust, such knowledge may nonetheless have the potential to give us pause before storming ahead with imposing policies only for experimental work.

Two aspects are relevant when discussing the potential problems of a literature. First, is there evidence that papers are suffering from questionable research practices? Second, to what extent is the literature "policing" itself, that is replicating important results and testing their robustness?

In Coffman and Niederle (2015) we discuss evidence by Brodeur, Lé, Sangnier, and Zylberberg (2016) whether *p*-hacking and questionable research practices are a substantial problem in applied economics and whether this problem is more pervasive in experimental economics. Analyzing all z-statistics in the American Economic Review, Journal of Political Economy, and Quarterly Journal of Economics between 2005 and 2011, they find, for paper using non-experimental data, evidence of questionable research practices.[72] In the absence of *p*-hacking, one would expect a perfectly smooth distribution of z-statistics, with perhaps one peak due to a threshold z-statistic for publication (implying that the literature suffers from publication bias). However, when authors use questionable research practices and engage in *p*-hacking to get *p*-values just below some desired thresholds, especially 0.05 (or a z-statistic just above 1.96), the distribution would have two peaks. This is because results that "just" fall short of a significance threshold are *p*-hacked to provide "nicer" results. This in turn generates "missing" z-statistics and hence a valley between the two peaks. Non-experimental applied papers present such a valley. However, papers using experimental data are largely single-peaked with a slight second bump, compared to the two sharp peaks from papers using non-experimental data.[73]

The second relevant aspect is whether the literature is likely to "correct" itself, that is, when a paper that is influential is based on results that are not robust, is the literature likely to find out? While this is hard to study, we can study the prevalence of replications.

Berry et al (2017) went through every single of the 70 empirical papers (excluding the P&P issue) published in 2010 in the *American Economic Review*. They checked every paper in the Web of Science (WoS) in June of 2016 citing one of those papers (restricting attention to papers published in a top 200 economics journal – using WoS impact factors) which lead to a sample of 1,558 citations. Of those 52 were replications (all but 4 without any author overlap with the original paper) with only 20 papers having a replication. A total of 42 of the 70 papers have any replication, extension or robustness check. Dividing the 70 empirical papers into fields (7 applied, 6 development, 13 Labor/IO, 19 behavioral and 25 Macro/international and Trade), just over half of the behavioral/experimental papers had at least one replication attempt, while for other fields this was between 12 (Macro/Intl/Trade) and 33 (development) percent. When we including extensions, once more behavioral experimental economics has the highest rate, well above 70 percent of papers having either a replication or extension, followed by development, with Macro/Intl/Trade having the lowest percentage, below 50%.

Berry et all (2017) show that as a science, economics seems healthy, as every single paper in their sample with 100 or more published citations had at least one replication. Of course, it could also just be that fields with more papers and citations replicate work at a higher rate.

Whether the replication and extension rates are a glass half full or half empty, one result is clear, however, namely that work in behavioral and experimental economics, as well as work in development seem to show the least sign of *p*-hacking and seem to be the most replicated compared to other applied fields in economics.

When it comes to the effects of various policies, I want to first re-iterate that even if issues 1-3 are perfectly solved, results from papers will need to be replicated and extended, tested for robustness, so that issue 4 and 5, which affect not only the paper but the whole literature, can be taken care of. In Coffman and Niederle (2015) we describe how papers can be classified into categories ranging from a unique field experiment involving very expensive and comprehensive data collections, likely testing hypotheses that have been brought up in the literature beforehand, like the Oregon health insurance study (Finkelstein et al. 2012), to papers that venture into unchartered waters and test novel hypotheses, that, potentially, could be studied by

[72] A z-statistic is a measure of how likely a result is due to chance rather than a true finding, where the higher the absolute value of the z-statistic, the lower the associated *p*-value.

[73] In fact, data from experiments or RCTs is the only cut of the data they provide where they find that the distribution of z-stats is largely single-peaked.

several economists at the same time, or economists could study other novel hypotheses with a similar ex ante chance to be true. In the first case, the work is very unlikely to be replicated again, so, have procedures in place to reduce issues 1-3 as much as possible won't affect issues 4-5. In the second case, not only are issues 1-3 a problem, but so are issues 4 and 5!

Before discussing the role of policies, a final reminder from Coffman and Niederle (2015) which heavily builds on the framework of Ioannidis (2005). We discuss the terms that affect the probability that a published, positive result is true, the "positive predictive value." Our estimate is built on five parameters. The first is the statistical significance threshold for a positive result (we used 0.05). The second parameter is the "power" of the study, which is 1 minus the "type II" error, the probability that a study will fail to detect an effect when an effect actually exists, we used 0.8. The third parameter is the expected probability of a hypothesis being true. The fourth parameter is the study bias, the probability that a study that would have been reported false without any bias is instead reported positive, basically putting a value on the chance that p-hacking can turn a "paper" around. The final parameter is the number of substitute studies that were (or could be) investigated, specifically, out of that number of studies, only the first positive one is reported, and others are either never investigated, or not finished and not reported.

For papers for which there are no substitute studies, like the Oregon health insurance study (Finkelstein et al. 2012), policies that restrict the study bias are extremely important. For example if the prior was a 50 percent chance that the hypothesis is correct, then, with a study bias of 25% the posterior is 75%, while is it 85% and 93% if we reduce study bias to 10% and 1%, respectively. In contrast, when there are 10 competing studies, the posteriors are only 51%, 56% and 71%, for study biases of 25%, 10% and 1%, respectively. If there are 25 studies, the corresponding numbers are 50%, 50% and 58%! That is, as the number of substitute studies increases, what really affects the chance the hypothesis is true after a positive result, is not so much the study bias, but rather the number of competing studies.

It is, however, not accurate to think of "substitute studies" as just the number of studies looking at the exact same hypothesis. For example, a researcher may have multiple distinct projects, each testing a different (though, for the sake of the argument) equally likely hypothesis, but then only decides to write up the first project with a positive result while letting others get filed away. For another example, a researcher may run (potentially informal) pilot studies to assess which hypothesis may be the most likely to yield a statistically significant result and only pursue the most promising pilot. Alternatively, the researcher may do this in their head (a poor researcher's version of pilots and an almost inevitable form akin to *g*-hacking). While ten such pre-tests (actual tests, pilots, or even thought of designs) are clearly not ten independent tests of the same hypothesis, it is also clear that they are not the same as just testing one hypothesis. Finally, it could be that 10 or 25 different economists test hypotheses with (for the sake of the argument) equally likely hypothesis, but only those with a positive effect finish their papers and publish them, then, once more, we are not at 10 or 25 independent tests of the same hypothesis, but clearly again far from just testing one hypothesis.

To summarize, even in the best-case scenario, with a very restrictive plan that avoids *p*-hacking as much as possible reducing it to maybe 1 percent (and ensuring the researcher does not write multiple pre-analysis plans), I am happy to be bold and claim that for almost all research the really damming effect comes from the number competing papers in a broad sense. In experimental economics this includes pre-testing and g-hacking.

Therefore, I feel that for work in experimental economics and laboratory experiments specifically, pre-registration and pre-analysis plans have arguably only a limited effect. There is another reason for this: When we look at a laboratory experiment, it is, in general, obvious what the paper wanted to do. Furthermore, the authors have, in general, only a very limited number of options on how to analyze the data. This may be somewhat different in field studies where there may be multiple variables to choose from when deciding whether to include them in the analysis and how, as field work, in general, operates in a richer setting than laboratory work. The latter often only has a very limited set of variables, making it difficult to have many meaningful different specifications to choose from. This is of course not always true, but, well, mostly it is. In general, I feel when I see the design of a laboratory experiment what individuals wanted to do, how I would analyze the data, and that is often what people do.

Is there ever a room for pre-registration? Yes, there is. This is especially the case if a researcher plans on running a specific subsample, or considers a specific subset of variables only, then, pre-registering this will allow the researcher to show that this is what they had in mind beforehand, rather than after playing with the data.

Are there costs to pre-analysis plans and pre-registrations? I worry that pre-analysis plans, are especially costly for young researchers, and may, for all researchers, increase the rate with which they pre-test their designs, and may lead to a tendency to use only minor deviations from existing designs, as then the analysis can be copied. Furthermore, strict pre-analysis plans may make researchers, and perhaps especially less experienced researchers more reluctant to head into the wilderness and test completely new hypotheses, where a rigid pre-analysis plan may be especially difficult to write.

I don’t see a real downside to pre-registration, I just think they may not be as useful as we may hope, since the file drawer problem is enormous, and, again, researchers may vary at what time they start writing such a plan, before or after the first pilot?

I do believe that in experimental economics, and most of economics, the much larger issue, which is an issue more for the literature perhaps than for a specific paper, are issues 4 and 5. We need to ensure that we have a lively and vibrant literature, where important results are routinely replicated, and tested for their robustness, taken to different environments and games. I worry that pre-analysis plans specifically, as they require the researcher to commit to analysis before having seen the data may stifle innovation, but also robustness tests and expansions that move far from the original paper.

My detailed opinions on pre-analysis plans are available to everyone (see Coffman, and Niederle, 2015, and Coffman, Niederle and Wilson, 2017), and I have, in a discussion in a roundtable of the ESA (Economic Science Association) meeting, posed the following question: “Do I trust a researcher’s results more when they (proudly) announce that they pre-registered their hypotheses and have a pre-analysis plan?” The answer I gave is “No,” and I still hold with this answer.

This is because I am aware of the many reasons beyond the manipulability of pre-analysis plans or the timing of pre-registrations I discussed above. We need replications anyway, not only to check the current result, but to understand its robustness. I think as Economists we very much underappreciate the value of replications (in a broad sense, that is including replications of the concept in different settings). If pre-registrations or pre-analysis plans further make us confident in the results of a given paper, I worry that this can backfire and make us underappreciate replications even more so.

That the worry that researchers believe that pre-registered papers are more "believable" is justified, is perhaps best exemplified by a paper in Nature: Protzko, J., Krosnick, J., Nelson, L. et al. (2024a) published a paper whose abstract contains:

> "This paper reports an investigation by four coordinated laboratories of the prospective replicability of 16 novel experimental findings using rigour-enhancing practices: confirmatory tests, large sample sizes, preregistration and methodological transparency. In contrast to past systematic replication efforts that reported replication rates averaging 50%, replication attempts here produced the expected effects with significance testing (P < 0.05) in 86% of attempts, slightly exceeding the maximum expected replicability based on observed effect sizes and sample sizes." [Note that these are largely results not from experimental economists.] [..] "This high replication rate justifies confidence in rigour-enhancing methods to increase the replicability of new discoveries."

As it happens, this paper is retracted! The following is part of the retraction notice

> "The Editors are retracting this article […].
>
> The concerns relate to lack of transparency and misstatement of the hypotheses and predictions the reported meta-study was designed to test; lack of preregistration for measures and analyses supporting the titular claim (against statements asserting preregistration in the published article); selection of outcome measures and analyses with knowledge of the data; and incomplete reporting of data and analyses.
>
> Post-publication peer review and editorial examination of materials made available by the authors upheld these concerns. As a result, the Editors no longer have confidence in the reliability of the findings and conclusions reported in this article."

Do I think pre-registrations and pre-analysis plans can sometimes be useful, yes, as I said, I do. This is the case when you want to replicate your own findings or have a specific analysis that relies on some complicated cuts of the data that may seem implausible ex post, or if you (but almost none of us does) run an experiment that cannot be replicated. However, for most papers, these arguments do not apply, and someone who then proudly engages in an activity I do not believe in (and doesn't just do it because some editors and referees require it), well, I will think less of that person, as it shows they may not understand the problems and limitations that come with such a pre-registration and such a pre-analysis plan, and what the real issues are in understanding the robustness of a literature. Finally, as the Nature paper showed, even individuals who are proponents of such policies seem to violate them without being explicit about it.

6. Conclusion

For a real conclusion, I want to end with a more positive message: For anyone designing or consuming experiments: A great experiment is almost like a performance! It is supposed to look simple, obvious and effortless. This does not mean that it was easy, simple or quick to get there. I hope this chapter provides you with some guidance along the way. Like the perfect simple example in a theory paper, a great experiment is the final product of often a long process. Do not be discouraged, first designs, like first drafts of a paper, are not always as beautiful and clean as final drafts. Find friends and colleagues who are happy to disagree with you and provide you with feedback. Co-authors are another great way to keep going and keep yourself honest about your design. Experiments are a lot of fun, and I keep learning new things from every experiment; this is, after all, why we run them!

References

Allen, Franklin, "Discovering personal probabilities when utility functions are unknown," *Management Science*, 1987, 33 (4), 542–544.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. Econometrica, 21(4):503–546.

Andreoni, James, "Cooperation in Public-Goods Experiments: Kindness or Confusion?," *The American Economic Review*, Vol. 85, No. 4 (Sep., 1995), pp. 891-904.

Andreoni, James, "Why Free Ride? Strategies and Learning in Public Goods Experiments." *Journal of Public Economics*, December 1988, 37(3), pp. 291-304.

Andreoni, James and Charles Sprenger, "Estimating time preferences from convex budgets," American Economic Review, 2012, Volume102, Issue 7, Pages 3333-3356.

Augenblick, Ned, Muriel Niederle and Charles Sprenger, "Working Over Time: Dynamic Inconsistency in Real Effort Tasks"*, Quarterly Journal of Economics*, August 2015, 130 (3): 1067-1115.

Azrieli, Yaron, Christopher P. Chambers, and Paul J. Healy (2018): "Incentives in Experiments: A Theoretical Analysis," Journal of Political Economy, 126(4), 1472–1503.

Azrieli, Yaron, Chambers, C.P. & Healy, P.J. Incentives in experiments with objective lotteries. *Exp Econ* **23**, 1–29 (2020).

Bajari, Patrick, and Ali Hortacsu, "Are Structural Estimates of Auction Models Reasonable? Evidence from Experimental Data," *Journal of Political Economy*, v. 113, Aug. 2005, 703-741.

Benoit, Jean-Pierre and Juan Dubra, "Apparent Overconfidence," *Econometrica*, 09, 2011, 79 (5), 1591–1625.

Berry, James, Lucas C. Coffman, Rania Gihleb, Douglas Hanley, and Alistair J Wilson, "Assessing the Rate of Replications in Economics,"*American Economic Review, Papers and Proceedings*, May 2017, 107(5): 27-31.

Bertrand, Marianne, Dean Karlin, Sendhil Mullainathan, Eldar Shafir & Jonathan Zinman, "What's Psychology Worth? A Field Experiment in the Consumer Credit Market," 2005, NBER Working Paper 11892

Binmore, Ken, Avner Shaked, and John Sutton, "Testing Noncooperative Bargaining Theory: A Preliminary Study," *American Economic Review*, 1985, 75, 11 78-80.

Blavatskyy, P., Ortmann, A., and Panchenko, V. (2022). On the experimental robustness of the allais paradox. American Economic Journal: Microeconomics, 14(1):143–63.

Blavatskyy, P., Panchenko, V., and Ortmann, A. (2023). How common is the common-ratio effect? Experimental Economics, pages 253—272.

Bolton, Gary E. and Rami Zwick, "Anonymity versus Punishment in Ultimatum Bargaining," Games and Economic Behavior, 1995, 10, 95-121.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2022. "Salience." Annual Review of Economics, 2022 14: 521-544.

Bouwmeester, Samantha et al. 2017. "Registered Replication Report: Rand, Greene, and Nowak (2012)." Perspectives on Psychological Science, 12(3): 527–542.

Brodeur Abel, Mathias Lé, Marc Sangnier and Yanos Zylberberg, "Star Wars: the Empirics Strike Back," *American Economic Journal: Applied Economics*, Jan. 2016, Vol. 8, Issue 1: p. 1-32.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek, "Gender, Competitiveness, and Career Choices," *Quarterly Journal of Economics,* August 2014, 129 (3): 1409-1447.

Buser, Thomas, Muriel Niederle and Hessel Oosterbeek, "Can Competitiveness Predict Education and Labor Market Outcomes? Evidence from Incentivized Choice and Survey Measures," forthcoming, *Review of Economics and Statistics*.

Camerer, Colin F. and Robin M. Hogarth, "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework," *Journal of Risk and Uncertainty*, 19, 7–42 (1999).

Capen, Edward C, Robert V Clapp, and William M Campbell (1971), "Competitive bidding in high-risk situations." *Journal of petroleum technology*, 23, 641–653.

Carpenter, Jeffrey, Eric Verhoogen and Stephen Burks, "The effect of stakes in distribution experiments," *Economics Letters*, Volume 86, Issue 3, 2005, Pages 393-398.

Chandrasekhar, Arun Gautham and Juan Pablo Xandri, "A note on payments in the lab for infinite horizon dynamic games with discounting," *Economic Theory,* 2023,**75**, 389–426.

Charness, Gary, Samek, Anya & van de Ven, Jeroen, "What is considered deception in experimental economics?" Experimental Economics 25, 385–412 (2022).

Coffman, Lucas C. and Muriel Niederle, "Pre-Analysis Plans Have Limited Upside Especially Where Replications Are Feasible," *Journal of Economic Perspectives*, Vol 29, No 3, Summer 2015, 29(3): 81-98.

Coffman, Lucas C., Muriel Niederle and Alistair J. Wilson, "A Proposal to Organize and Promote Replications," *American Economic Review, Papers & Proceedings,* vol. 107, no. 5, May 2017, 41-45.

Cohen, Shani and Shengwu Li, "Sequential Cursed Equilibrium," working paper, 2024.

Conlon, John J., "Attention, Information, and Persuasion," working paper, 2024.

Costa-Gomes, Miguel, A., and Vincent P. Crawford. 2006. "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study." *American Economic Review*, 96 (5): 1737–1768**.**

Crawford, Vincent P and Nagore Iriberri (2007), "Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions?" *Econometrica*, 75, 1721–1770.

Crosetto, P., Filippin, A. A theoretical and experimental appraisal of four risk elicitation methods. *Experimantal Economics* **19**, 613–641 (2016).

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang. 2007. "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness." *Economic Theory*, 33: 67–80.

Dean, Joshua, Christine Exley, Muriel Niederle and Heather Sarsons, "Measuring Gender Attitudes toward Men and Women," working paper, 2024.

DeSousa, José and Muriel Niederle, "Trickle-Down Effects of Affirmative Action: A Case Study in France," August, 2023.

Eckel, C. C., & Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, *23*(4), 281–295.

Einav, Liran, Amy Finkelstein, Iuliana Pascu and Mark R. Cullen, "How General Are Risk Preferences? Choices under Uncertainty in Different Domains," American Economic Review vol. 102, no. 6, October 2012, 2606–38

Enke, Ben, "The Cognitive Turn in Behavioral Economics," working paper, 2024.

Enke, Ben, and Thomas Graeber, "Cognitive Uncertainty," *Quarterly Journal of Economics, 2023, vol. 138 (4), pp. 2021-2067.*

Erev, Ido, Alvin E. Roth and Robert Slonim, "Minimax across a population of games" J Econ Sci Assoc (2016) 2:144–156.

Exley, Christine, "Incentives for Prosocial Behavior: The Role of Reputations," *Management Science*, 2017.

Exley, Christine, L and Judd Kessler, "Information Avoidance and Image Concerns," *Economic Journal*, November 2023, 133 (656): 3153-3168.

Eyster, Erik and Matthew Rabin (2005), "Cursed equilibrium?" *Econometrica*, 73, 1623–1672.

Finkelstein, Amy, Sarah Taubman, Bill Wright, Mira Bernstein, Jonathan Gruber, Joseph P. Newhouse, Heidi Allen, Katherine Baicker, and the Oregon Health Study Group. 2012. "The Oregon Health Insurance Experiment: Evidence from the First Year." *Quarterly Journal of Economics* 127(3): 1057–1106.

Friedman, Daniel, R. Mark Isaac, Duncan James and Shyam Sunder, "Risky Curves: On the empirical failure of expected utility," Routledge, 2013. ISBN No. 978-0-415-63610-0.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini, "Performance in Competitive Environments: Gender Differences," *Quarterly Journal of Economics*, CXVIII, August 2003, 1049 – 1074.

Gneezy, Uri and Potters, Jan (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, *112*(2), 631–645.

Gneezy, Uri and Aldo Rustichini, "Pay enough of don't pay at all," *The Quarterly Journal of Economics*, August 2000, 791-810,

Goscinny, René, Albert Uderzo, "Asterix le Gaulois," Dargaud, 1961.

Goscinny, René, Albert Uderzo and Anthea Bell, "Asterix the Gaul," series: Asterix: Orion Paperback, Hodder & Stoughton, 1969.

Grether DM (1992) Testing Bayes rule and the representativenessheuristic: Some experimental evidence. J. Econom. Behav. Organ.17(1):31–57

Griggs, R. A., & Cox, J. R. (1982). "The elusive thematic materials effect in Wason's selection task" *British Journal of Psychology*,73(3), 407-420.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze, "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization*, 1982, 3, 367-88.

Handel, Benjamin and Joshua Schwartzstein (2018), "Frictions or mental gaps: what's behind the information we (don't) use and when do we care?" Journal of Economic Perspectives, 32, 155–178.

Harrison, Glenn W., "Theory and Misbehavior of First-Price Auctions," American Economic Review, September 1989, 79, 749-62

Holt CA (1986) Preference reversals and the independence axiom. Amer. Econom. Rev. 76(3):508–515.

Holt, C., & Laury, S. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.

Huang, Jennie, Judd B. Kessler, and Muriel Niederle, "Fairness has less impact when agents are less informed," *Experimental Economics* (2024) 27:155–174.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124.

Jain, Ritesh and Kirby Nielsen, "A Systematic Test of the Independence Axiom Near Certainty," working paper, 2024.

John Leslie K., George Loewenstein, Drazen Prelec. "Measuring the prevalence of questionable research practices with incentives for truth telling." Psychol Sci. 2012 May 1;23(5):524-32.

Kagel, John H. and Dan Levin, "The winner's curse and public information in common value auctions." *American Economic Review*, 1986, 894–920.

Kagel, John H. and Dan Levin, "Independent Private Value Auctions: Bidder Behaviour in First-, Second- and Third-Price Auctions with Varying Numbers of Bidders," *The Economic Journal*, Vol. 103, No. 419. (Jul., 1993), pp. 868-879.

Kagel, John H and Dan Levin (2002), "Common value auctions and the winner's curse." 2002

Kagel, John H. and Alvin E. Roth, "The Dynamics of Reorganization in Matching Markets: A Laboratory Experiment Motivated by a Natural Experiment," *The Quarterly Journal of Economics*, Vol. 115, No. 1 (Feb., 2000), pp. 201-235.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2):263–292.

Karni E (2009) A mechanism for eliciting probabilities. Econometrica77(2):603–606.

Kessler, Judd B., Hannu Kivimaki, Ashley Litwin, and Muriel Niederle, "Please Take a Minute: How Prosocial Choices Change with Deliberation," working paper 2024.

Kreps, David M, Paul Milgrom, John Roberts and Robert Wilson, "Rational cooperation in the finitely repeated prisoners' dilemma, Journal of Economic Theory, Volume 27, Issue 2, 1982, Pages 245-252.

Liberman, Varda, Samuels, Steven M., & Ross, Lee (2004). "The Name of the Game: Predictive Power of Reputations versus Situational Labels in Determining Prisoner's Dilemma Game Moves." *Personality and Social Psychology Bulletin*, 30(9), 1175-1185.

Loewenstein, George, and Zachary Wojtowicz, "The Economics of Attention," working paper, 2023.

Martinez-Marquina, Alejandro, Muriel Niederle and Emanuel Vespa, "Failures in Contingent Reasoning: The Role of Uncertainty," *American Economic Review*, vol. 109, no. 10, October 2019, 3437-74.

Marwell, Gerald and Ruth E. Ames, 1981, "Economists free ride: Does anyone else? Experiments on the provision of public goods," IV, *Journal of Public Economics* 15, 2955310.

McGranaghan, Christina, Kirby Nielsen, Ted O'Donoghue, Jason Somerville, and Charles D. Sprenger (2024): "Distinguishing Common Ratio Preferences from Common Ratio Effects Using Paired Valuation Tasks," *American Economic Review*, 114(2), 307–347.

McGranaghan, Christina and Nielsen, Kirby and O'Donoghue, Ted and Somerville, Jason and Sprenger, Charles D., "Connecting Common Ratio and Common Consequence Preferences," working paper, 2024.

Mobius, Markus M., Muriel Niederle, Paul Niehaus and Tanya Rosenblat, "Managing Self-Confidence: Theory and Experimental Evidence," *Management Science*, 68(11), 2022, 7793-7817.

Moore, D. A., and Healy, P. J. (2008). "The trouble with overconfidence," *Psychological Review*, 115(2), 502–517.

Murnighan, J.Keith and Alvin E. Roth, "Expecting continued play in prisoner's dilemma games a test of several models," Journal of Conflict Resolution 27(2), 279–300 (1983).

Nagel, Rosemarie. "Unraveling in Guessing Games: An Experimental Study." *The American Economic Review*, vol. 85, no. 5, 1995, pp. 1313–26.

Nagel, Lea, Muriel Niederle and Emanuel Vespa, "Decomposing the Winner's Curse," working paper, 2025.

Nash, John F. The bargaining problem. Econometrica, 1950, 28, 155-162.

Ngangoue, M Kathleen and Andrew Schotter (2023), "The common-probability auction puzzle?" *American Economic Review*, 113, 1572–1599.

Niederle, Muriel "Intelligent Design. The Relationship of Economic Theory to Experiments: Treatment driven Experiments" *Handbook of Experimental Economic Methodology*, edited by Guillaume R. Fréchette and Andrew Schotter, Oxford University Press, February 2015, 104-131.

Niederle, Muriel, "Gender" Handbook of Experimental Economics, second edition, Eds. John Kagel and Alvin E. Roth, Princeton University Press, 2016, pp 481-553.

Niederle, Muriel and Emanuel Vespa, "Decisions under Uncertainty: Risk Preferences or Complexity?", working paper, 2019.

Niederle, Muriel and Emanuel Vespa, "Cognitive Reasoning: Failures of Contingent Thinking," *Annual Review of Economics, 2023, Vol. 15, 307–328*

Niederle, Muriel, and Lise Vesterlund, "Do Women Shy Away from Competition? Do Men Compete too Much?," *Quarterly Journal of Economics,* August 2007, Vol. 122, No. 3, 1067-1101.

O'Neill, Barry, "Nonmetric test of the minimax theory of two-person zerosum games," *Proc. Natl. Acad. Sci.* Vol. 84, pp. 2106-2109, April 1987

Oprea, Ryan, "Decisions Under Risk are Decisions Under Complexity," forthcoming at the American Economic Review, 2024.

Protzko, J., Krosnick, J., Nelson, L. et al. RETRACTED ARTICLE: High replicability of newly discovered social-behavioural findings is achievable. Nat Hum Behav 8, 311–319 (2024a). https://doi.org/10.1038/s41562-023-01749-9

Protzko, J., Krosnick, J., Nelson, L. et al. "Retraction Note: High replicability of newly discovered social-behavioural findings is achievable." Nat Hum Behav 8, 2067 (2024b). https://doi.org/10.1038/s41562-024-01997-3

Puri, Indira, "Simplicity and Risk," *Journal of Finance*, Forthcoming. May 2024

Recalde, Maria, Arno Riedl and Lise Vesterlund, "Error Prone Inference from Response Time: The Case of Intuitive Generosity," *Journal of Public Economics*, 160, 2018, pp. 132-147.

Roth, Alvin E. and Michael W. K. Malouf, "Game-Theoretic Models and the Role of Information in Bargaining," Psychological Review, 1979, Vol. 86, No. 6, 574-594.

Roth, Alvin E., and J. Keith Murnighan. 1978. “Equilibrium Behavior and Repeated Play of the Prisoner’s Dilemma.” Journal of Mathematical Psychology 17 (2): 189–98.

Roth, Alvin E. and J. Keith Murnighan, “The Role of Information in Bargaining: An Experimental Study,” *Econometrica*, Vol. 50, No. 5(Sep), 1982), 1123-1142.

Roth, Alvin E, “Let's Keep the Con Out of Experimental Econ.: A Methodological Note,” *Empirical Economics* (Special Issue on Experimental Economics), 1994, 19, 279-289.

Roth, Alvin E, "A Natural Experiment in the Organization of Entry Level Labor Markets: Regional Markets for New Physicians and Surgeons in the U. K.," *American Economic Review*, LXXXI (June 1991), 415-440.

Roth, Alvin E. “The origins, history, and design of the resident match,” *JAMA. Journal of the American Medical Association*, vol. 289, No. 7, February 19, 2003, 909-912.

Roth, A.E. and E. Peranson, "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design,” *American Economic Review*, 89, 4, September, 1999, 748-780.

Segal, Carmit, “Working When No One Is Watching: Motivation, Test Scores, and Economic Success,” *Management Science*, Vol. 58, No. 8, August 2012, pp. 1438–1457.

Segal, Uzi, “Two-Stage Lotteries without the Reduction Axiom,” Econometrica, Vol. 58, No. 2 Mar., 1990, pp. 349-377.

Sherstyuk, Katerina, Nori Tarui and Tatsuyoshi Saijo, “Payment schemes in infinite-horizon experimental games,” *Experimental Economics* 16 (2013), 125–153.

Simonsohn U, Nelson LD, Simmons JP. p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. Perspect Psychol Sci. 2014.

Smith, Vernon L., 1976, “Experimental economics: induced value theory,” *American Economic Review,* 66, 274-279. TO CITE SOMEWHERE

Strunk, William Jr. and Elwyn Brooks White, “The Elements of Style.” S.l.: Longman, 2000. ISBN 0-205-30902-X (paperback).

Wason, Peter C. (1966). Reasoning. In B. Foss (Ed.), New horizons in psychology (pp. 135-151). Harmondsworth: Penguin Books.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. Quarterly Journal of Experimental Psychology, 23, 63-71.