Non est Disputandum de Generalizability?
A Glimpse into The External Validity Trial

John A. List
University of Chicago[1]

4 July 2020

Abstract

While empirical economics has made important strides over the past half century, there is a recent attack that threatens the foundations of the empirical approach in economics: external validity. Certain dogmatic arguments are not new, yet in some circles the generalizability question is beyond dispute, rendering empirical work as a passive enterprise based on frivolity. Such arguments serve to caution even the staunchest empirical advocates from even starting an empirical inquiry in a novel setting. In its simplest form, questions of external validity revolve around whether the results of the received study can be generalized to different people, situations, stimuli, and time periods. This study clarifies and places the external validity crisis into perspective by taking a unique glimpse into the grandest of trials: The External Validity Trial. A key outcome of the proceedings is an Author Onus Probandi, which outlines four key areas that every study should report to address external validity. Such an evaluative approach properly rewards empirical advances and justly recognizes inherent empirical limitations.

[1]Some of the people represented in this trial are creations of the mind, as are the phrases and statements that I attribute to the real people in this trial. Specifically, these words do not represent actual statements from the real people, rather these are words completely of my creation. I have done my best to maintain historical accuracy. Thanks to incredible research support from Ariel Listo, Lina Ramirez, Haruka Uchida, and Atom Vayalinkal. Min Sok Lee provided quite helpful information for this study. Working with many coauthors as well as discussions/debates over the years with colleagues helped me discover my voice and whereabouts on this issue. I trust that they will not agree with all of the views expressed in this satirical study. Omar Al-Ubaydli, Alec Brandon, Cody Cook, Aart de Zeeuw, Rebecca Diamond, Seda Ertac, Basil Halperin, Justin Holz, Robert Metcalfe, Magne Mogstad, David Novgorodsky, Michael Price, Matthias Rodemeier, Kerry Smith, and Karen Ye provided quite helpful comments on earlier drafts.

Induction or generalization is never fully justified logically
~Hume's truism

Bailiff Barney Fife: Order, order, all rise for the honorable Judge Learner of the Court of Kanga Roo.

Judge Ima Learner: Thank you Bailiff Fife. When these proceedings started, my primary goal was to gather enough information to rule on the validity of the current empirical approach in economics; and, if found to be logically sound, to put forward an Author Onus Probandi as a constructive, theoretically-driven, approach to external validity. I plan to provide that checklist at the end of today's proceedings. Our proceedings started 1968 days ago, when, against Kanga precedent, I allowed the defense to begin with their witnesses. We have now welcomed more than 777 witnesses into our Supreme International Court on Empirical Methods. Scholars, practitioners, scientists, entrepreneurs, politicians, poets, philosophers, and key experts have traveled here to the far reaches of the Australian Outback to provide sage advice and broad wisdom. The amount of information has been breathtaking. For this purpose, before we welcome our last set of witnesses today, we need a succinct summary detailing for the Court the major points raised to date. Runner Clifton Hillegass, the floor is yours.

Court Reporter Clifton Hillegass: Thank you Judge Learner. While it is never easy to convey succinctly the key points of a debate, this dispute has crystallized in a manner that leaves no middle ground. The Prosecution, led by Mr. Naiv Ete, argues that all empirical work in economics must pass a set of necessary external validity conditions before being published in academic journals or used by policymakers. To date, in this courtroom no empirical work has passed his conditions, effectively rendering the question of generalizability beyond dispute, or as Livius Andronicus reminded us, Non est Disputandum de Generalizability. Ms. Minerva, Lead Defense, has argued that this line of reasoning leaves only theoretical exercises and thought experiments to advance science and guide policymaking, an approach that she fears will return us to the dark ages.

To provide an understanding of the terminology, Professor Campbell testified on the language of mainstream empiricism, which is rooted in the taxonomy of threats he and Stanley introduced in 1966, later extended by his work with Cook in 1979 (Campbell and Stanley, 1966;

Campbell and Cook, 1979). The four types of validity that the literature has focused on are statistical conclusion validity, internal validity, construct validity, and external validity. For our purposes, the literature has come to use the terms "external validity" and "generalizability" interchangeably, as external validity pertains to the question of generalizability: to what populations of situations, people, and temporal settings can this effect be generalized? It is commonly argued that internal validity—inferences about whether observed covariation between Y and X reflects a true causal relationship from X to Y—is a prerequisite for external validity because study results that deviate from the true effect due to systematic error lack the basis for generalizability.[2]

In this manner, another way to characterize generalizability of results is to describe their portability. For instance, will a study's conclusions remain valid if transferred to others in the experimental population from which the researcher drew her sample? What about to populations and situations not included in the study? For example, from the laboratory to naturally-occurring markets?[3] For organizational-level work, the generalization question revolves around the notion of: we found a large treatment effect with Amazon.com employees but does the same relationship hold with Uber, Google, Alibaba, Ford Motors, Citibank, and McDonald's employees? Further, generalizability refers to one's ability to transfer results to a different time period: Henry Ford's efficiency wage idea worked with Ford employees in 1914 but would it work for Ford's employees today? Perry pre-school worked for children of the 60s, does it work now?

With these concepts in hand, the Defense attempted to have the entire case dismissed based on the "Rome wasn't built in one day" argument. Ms. Minerva brought forward Francis Bacon, whose cogent argument in the *Novum Organum* demands researchers slowly build an essential base of knowledge from the ground up, one experiment at a time. Likewise, as Professor Marie Curie noted, the "successful" sciences, such as the two fields in which she earned a Nobel prize,

[2] One can trace internal validity back to at least Robert Koch, whose postulates for proof of causation in biological studies helped form the basis for judging and rating internal validity (see Gradmann, 2004). Hill (1965) used the postulates as the basis for his classical five criteria for health research. Campbell and Stanley's (1963) widely cited "threats to internal validity" remain the modern benchmark.

[3] Some have argued that external validity is purely a sample size problem. This is partly true. Where external validity refers to generalizing to the rest of the same population from which a sample is taken, increasing the size of the sample does improve inference. However, where external validity refers to a population of different situations or people different from the populations from which the original research sample was drawn, increasing the size of the sample of the original study would not necessarily improve the portability to these different populations (unless these new situations or people themselves are sampled when the sample size is increased).

tend to make strides one step at a time with a blind eye toward representativeness concerns, instead leveraging the beauty of the laboratory setting because the artificiality it provides is essential for the separation of variables fundamental to their theories. In this way, empirical work is valuable regardless of its external validity because it both tests theory and augments scientific knowledge, even if only incrementally. Both the Baconian approach and Professor Curie's view of scientific progress was that it was never swift, and Professor Curie warned us to "never fear perfection in a study, because you will never reach it." Mr. Naiv Ete likened this dismissal challenge to a "Kuhnian paradigm shift," and our High Court continued with arguments.

Professor Curie's statements naturally motivated a discussion of how extant theory can aid our understanding of generalizability. Professors Levitt and Al-Ubaydli testified about their work, with the latter focusing "On the generalizability of experimental results in economics" (2013), which extends the all causes model to a more continuous form where researchers have priors about causal effects and update them based on data. Interestingly, beyond the typical comparisons across domains, Professor Al-Ubaydli emphasized that it is important to consider how selection rules differ between the experiment versus the target setting of interest, whether the experiment places artificial restrictions on individual's decision/choice sets, and the time horizon of choices (see also Levitt and List, 2007a; 2007b). His model naturally arose from Mill's assumption of the lawfulness of nature in that it is based on the "distance" between time, space, population, and the decision environment between the study setting and the setting of ultimate interest.[4] Generalization in this case revolves around the degree of "stickiness" of nature: the model assumes that as two decisions become closer in distance, the congruency in response effects will heighten because the two environments tend to follow similar laws.

One approach that provides empirical guidance into this type of theoretical exercise was discussed in Professor Guido Imbens testimony, who testified about his 2005 work with Joseph Hotz and Julie Mortimer (Hotz et al., 2005). Their exploration pertained to the transferability of insights across domains. They operationalized the idea by examining whether they could predict

[4] Distance in this sense is in the spirit of what MacDougall recognized in 1922 and what Brunswik (1952; 1956) pioneered with his ideas pertaining to representative design, which advocated replacing systematic design with a design leveraging situations representative of the subject's ecology, or habitat. Representative design and a related concept, ecological validity, have been abused and redefined over the years; the concept of ecological validity, as originally defined, refers to the potential utility of various cues for organisms in their natural habitat. Thus, in cases where the target setting contains important cues for decisionmakers, a representative design will contain those same cues and their intercorrelations.

the effect of a training program in one location given information they learned on program effects from four very different locations. Their work is consonant with the theory in that they found that even if the original result was not measured with populations and in situations of ultimate interest, the analyst could in principle use a statistical approach whereby appropriate adjustments are made to account for differences across environments. Their case study provides some optimism that differences in populations can be adjusted to yield insights across settings, suggesting that there is information from measuring effects outside of the ultimate domain of interest.

Ultimately, what the literature teaches us is that our attempts to generalize are basically guesses at laws and examining whether these generalizations are valid. In the course of scientific discovery itself, we learn about the justification of generalizing by the accumulation of our experiences, but this is not a logical generalization deducible from the details of the original experiment. In such cases, we can use theory to guide us about the "generalizability guesses" as to yet unproven laws. Importantly, external validity can be viewed as what is valid in an experiment depends on the theoretical framework that is being used to draw inferences from the observed behavior in the experiment.

We have learned that the history of science is replete with illustrations of incomplete theories being tested. Consider Heinrich Hertz's experimental work in electromagnetic theory. Should he have had a model that included all background factors that could affect radio waves? Color of his coat, different types of measurement instruments, color of paint on the lab walls, size of the lab? His single best source of guidance about these choices, electromagnetic theory, was silent on the endless factors that could matter. The fact that size of lab mattered in the end gives us a useful analogy of how science progresses. If we have a theory that (implicitly) says that lab size does not affect radio waves, then any experiment that ignores lab size is valid from the perspective of that theory and is generalizable across labs of different size. In this spirit, one cannot begin to identify what factors make an experiment valid in an external sense without priors from a theoretical framework (see Harrison and List, 2004). Rome cannot be built in one day; Hertz missed key background factors that matter but his contribution is not diminished in the least.

Finally, as the sciences have matured, Professor Ioannides argued that greater weight should be placed on evidence based on multiple studies (see, e.g., Ioannides, 2005; Maniadis et al., 2014; Dreber et al., 2015). While he applauded the recent replication movement in the social sciences, the Prosecutor noted that step was largely due to the power of replicating results in

identical populations and situations, rather than representing a movement to exploring how results might vary across different populations and situations as well as when they are scaled. In this sense, scholars on both sides of the debate agree that there are gaps in our understanding about generalizability, both empirically and theoretically. The great irony, however, is that many scholars lament the fact that evidence-based practice has had difficulty gaining widespread traction while simultaneously ignoring the scientific gaps in our knowledge of generalizability and scaling.

Judge Ima Learner: Many thanks for an excellent set of "Cliff notes" Runner Hillegass. I suspect that we have been at this quite long enough. We must finish this case today with the remainder of the prosecution's witnesses and closing arguments. We must determine today if the empirical approach should continue within economics and the broader sciences, and if so in what form. Ms. Minerva and Mr. Ete, can you both meet my temporal demands?

Defender Minerva: Yes, the Defense can meet that request Judge Learner.

Prosecutor Naiv Ete: Judge Learner, we can meet that request. As you know this trial has gone very well for the State of Empirical Disbelief. We trust that we have shown the nonsensical nature of the empirical approach in economics through the very scientific legs it stands. Today we will seal this case by traversing centuries of empirical scholars with a focus on more recent economists, picking apart their results one by one. The State calls Louis Pasteur.

Louis Pasteur: I solemnly and sincerely declare and affirm that the evidence I shall give will be the truth, the whole truth, and nothing but the truth.

Prosecutor Naiv Ete: Thanks Mr. Pasteur. Can you tell me about your experimental origins?

Louis Pasteur: I was first introduced to the empirical approach in the 1840s as a graduate student at École Normale Supérieure (ENS), where I fell in love with laboratory work in Physics and Chemistry. Later, in 1877, while I was the director of ENS' laboratory of physiological chemistry, I was asked by one of my mentors to study the mysterious deaths of silkworms in south France,

which were devastating the silk industry in that region. My work on this problem gave me a basic understanding of contagion and stimulated my interest in contagious diseases.

In an exciting career moment, on October 30th, 1878, I received from Henry Toussaint, a Professor at the Veterinary School of Toulouse, a strain of bacteria that caused chicken cholera. I cultivated the chicken cholera microbe in chicken broth and began to perform experiments on chickens using these cultures. That next July, I instructed my assistant Charles Chamberland to inoculate a clutch of hens with cultures of the bacterium while I was away on vacation. Charles, however, failed to do this and went away on vacation himself, leaving the cultures in the lab. When he returned a few weeks later and inoculated the hens, he found that the (now old) cultures did not kill the hens, and instead left them immune to chicken cholera.

Further tests revealed that the bacteria had become weakened by exposure to oxygen in the air during Charles' vacation, and that they retained their capacity to protect hens from chicken cholera while losing their ability to cause the disease. We found that cultures aged for a few months formed what we would today call a vaccine - killing no chickens, while still protecting them from chicken cholera. My success with chicken cholera led me to further experiments seeking vaccines for other diseases – whereby I successfully developed vaccines for anthrax, swine erysipelas, and rabies (this passage draws heavily from Bazin (2011)).

Prosecutor Naiv Ete: Very good. Let's move to another one of your farm animals. Can you tell me about this famous sheep study of yours?

Louis Pasteur: Sure. By 1881 there were many skeptics of my germ theory of disease, perhaps none more vocal than a veterinarian named Monsieur H. Rossignol, who challenged me to test my theory in public by vaccinating sheep on his farm at Pouilly-Le-Fort, a small village outside of Paris. Given the public nature of his attack I was left no choice but to accept the challenge, even though it was risky because to my knowledge no vaccine had ever been tested outside the laboratory. On May 5, 1882, 25 sheep at Rossignol's farm were inoculated for "charbon," a disease today you call "anthrax". For the historians in the room, charbon is an ancient disease reported by Homer and Hippocrates as extremely deadly, known to kill thousands of animals each year back then. On May 17, 1882, these 25 sheep received another 'protective injection'. Beyond the 25 inoculated sheep, another 25 received no vaccine.

On May 31, all 50 sheep were injected with a culture of very virulent charbon. For me to be declared the winner, every control sheep had to die and every vaccinated sheep had to live. Given its importance and novelty of the event the publicity was intense. Reporters scribed daily reports for newspapers all around France; the London Times had a reporter dispatched to the farm to provide daily bulletins back to London. I was interviewed almost constantly it seemed—I had little chance to do my science with interviewers, press, onlookers, and gawkers constantly watching my every move. I was swooned over like Lola Montez, William Arden, or a modern day Kardashian!

Within 2 days, a group of farmers, veterinarians, pharmacists, and agriculture officials gathered at Rossignol's farm to observe the results of my field experiment. The experiment proved to be an overwhelming confirmation of my theory: 2 days after inoculation, every one of the 25 control sheep was dead whereas all 25 vaccinated sheep remained alive and well.

Prosecutor Naiv Ete: Very impressive victory for your past self, Dean Pasteur. Can you describe why you believe this set of findings is important?

Louis Pasteur: Well, beyond revealing the truth behind my theory and showing that I was the preeminent scientist of the day, I illustrated the power of the scientific method. Indeed, more narrowly, my results showed not only how we can protect animals from the virulent strain of the anthrax microbe, but also it proved the germ theory of disease. Protection against anthrax helped to establish vaccines against smallpox and other diseases. From my work emerged the disciplines of immunology and bacteriology, which eventually led to vaccination of millions of people and prevention of many diseases.

Prosecutor Naiv Ete: Hmmm…this all sounds impressive, but….may I ask, what color were your sheep?

Louis Pasteur: Well, I cannot really remember that was a long time ago, but I believe they were all white, yes, all 50 of them were white to my best knowledge.

Prosecutor Naiv Ete: *Aha, Dean Pasteur, some are born great, some achieve greatness, and some have greatness thrust upon them.* It seems that in this case the celebrity cast upon you is for utter nonsense. There are a lot of black sheep milling around this earth. And, we know that the sheep who participated in your experiment are not representative of the ovine family in the animal kingdom. Influenced by the ideal of the French Revolution (liberté, equalité, fraternité), your experimental subjects were given opportunities seldom available to their distant relatives, the guinea pigs, for example. This alone could lead to different results.

Louis Pasteur: Yes, but the scientific theory does not rely on hair color of the species, so ignoring hair color is valid from the perspective of my theory…

Prosecutor Naiv Ete: Enough, we have heard enough about the speculation of your theories Dean Pasteur. This study clearly has no external validity. Your sheep are not representative of other sheep or farm animals, much less the species of most interest, humans. Dean Pasteur, have fun on Green Acres, or wherever you are off to this afternoon. Judge Learner, I am done here.

Judge Ima Learner: Defense, would you like to cross this witness?

Defender Minerva: Beautiful early example of a courageous field experiment showing external validity, Professor Pasteur. The Defense thanks you for your enduring contributions to science and broader humanity. We all live healthier lives because of your work, which set the basis for how we attack the most recent health crisis to humanity: COVID-19. No questions from the Defense, Judge Learner.

Judge Ima Learner: Professor Pasteur, you are dismissed. The court thanks you for your time. Please call your next witness Mr. Ete.

Prosecutor Naiv Ete: The State calls Sir Ronald Fisher.

Ronald Fisher: I solemnly and sincerely declare and affirm that the evidence I shall give will be the truth, the whole truth, and nothing but the truth.

Prosecutor Naiv Ete: Thank you Mr. Fisher. Can you tell me about your experimental origins?

Ronald Fisher: My early passion was evolutionary theory and I gravitated to astronomy and physics in my schooling. I taught mathematics and physics from 1914-1919. In 1919, I became a statistician for the City of London, when I accepted an offer to work at the Rothamsted Experimental Station as a statistician. There, I analyzed the station's immense experimental data and developed several statistical techniques to help with the empirical analysis. Working at Rothamsted spurred my interest in the design of experiments, and I began thinking about how to improve the validity and efficiency of field experiments (these insights and those below are drawn from Box (1980), who also provides a more patient overview of Fisher).

Prosecutor Naiv Ete: Very good. Can you now tell me about these famous manure studies of yours?

Ronald Fisher: Sure. While at Rothamsted, I helped analyze data from experiments seeking to determine the response to manure spread across varieties of potatoes. A plot of land that had historically been treated uniformly with manure was divided into two equal parts, one part for farmyard manure and the other half for non-farmyard manure. Each part was then split into 36 patches, and a variety of potato was grown on each patch.

There were 12 potato varieties in all, and each variety was grown on 6 patches total, with 3 scattered over each half. Each patch was divided into three lines, one of which received, in addition to the corresponding variety of manure, a basal dressing only, containing no potash, while the other two received additional dressings of sulphate and chloride of potash respectively.

We found that while there is clearly significant variation in yield across manure treatment, and across varieties of potato, there is no significant variation in the response of different varieties to manure (see Fisher and Mackenzie (1923)).

Prosecutor Naiv Ete: Can you describe why you believe this set of findings to hold import?

Ronald Fisher: These findings, narrowly interpreted, are important because they helped dispel the notion, prevalent at the time, that varieties of cultivated plants differed not only in their suitability to different climatic and soil conditions, but also in their response to different manures. We randomized the variety of potato grown on each patch, and then tested each kind of dressing on every patch. In this way, we were able to separate variations in yield that were driven by differences between potato varieties, or by differences in manurial treatment, from variations in yield that were driven by variations in the response of different varieties to manurial treatment. As a result, we were able to show conclusively that different varieties of potato do not respond differently to types of manurial treatment, allowing farmers to make better decisions about which varieties of a plant to grow, and which varieties of manurial treatment to use.

Prosecutor Naiv Ete: And may I ask….what types of manure did you compare and where and when?

Ronald Fisher: Well, they were all either dung or green manure, some of which were mixed with different varieties of Potash and we did it during one crop cycle in England.

Prosecutor Naiv Ete: *Aha, fair is foul, and foul is fair: Hover through the fog and filthy air Sir Fisher*. As an agronomy minor in college at the University of Wisconsin at Sun Prairie (UWSP), I learned that the returns to manure type critically depend on time and place, and Sir Ronald Fisher you have one time and one small area in the UK. My learnings at UWSP therefore lead me to the fact that we learn nothing from your study. You didn't sample the correct times and places and your stimuli is barely used today in the form that you tested; this leads to an obvious conclusion: you have no external validity.

Ronald Fisher: Wait one moment Sir, that isn't correct because regardless of any particular study's results, historically Rothamsted is the birthplace for modern data analysis as my work developed the core concepts for experimental design and generation of data that are still taught today all over the world. Therefore, this study puts on full display the power of…..

Prosecutor Naiv Ete: Enough, enough already Mr. Fisher. You should find yourself ashamed, your research, together with Bacon, Pasteur, and Galileo's work, have been instrumental in getting us into this whole mess in the first place. We have had enough of your bullshit experiments!

Judge Ima Learner: Order, order Mr. Ete, I will not accept that type of language in my courtroom. Defense, would you like to cross this witness?

Defender Minerva: Quite an interesting example of your pioneering statistical design in a landmark set of field experiments, Professor Fisher. Experimentalists today continue to rely on your pathbreaking "tripod" insights, and for this it is a pleasure to be in your company. Your work published in 1925 was a true pleasure and remains a hallmark today (Fisher, 1925). No questions from the Defense, Judge Learner.

Judge Ima Learner: Sir Fisher, you are dismissed. The court thanks you for your time. Please call your next witness Mr. Ete, and please be on your best behavior.

Prosecutor Naiv Ete: The State calls Professor Vernon Smith, Professor from Chapman University.

Vernon Smith: I solemnly and sincerely declare and affirm that the evidence I shall give will be the truth, the whole truth, and nothing but the truth.

Prosecutor Naiv Ete: Thank you Professor Smith. Can you tell me about your experimental origins?

Vernon Smith: My story really starts in Wichita, Kansas where I was born into an inquisitive family. My tools were strongly technical as a lad, as I received my bachelor's degree in electrical engineering from Caltech in 1949. I returned to Kansas for a Master's degree in economics in 1952 and then went to Harvard for a PhD in 1955. My days at Harvard were foundational because they set the wheels in motion for what I would take as a life-long pursuit: experimentation.

In economics, I was first introduced to the experimental approach when I was an undergraduate student subject in Professor Chamberlin's laboratory experiment on market equilibria in the 1940s (Chamberlin, 1948). He used a decentralized auction approach and he found that in his multi-lateral exchange markets too many trades took place and the realized prices were too low according to standard economic theory. This instantly fascinated me because I was the "main course" of the experiment, acting as a subject in Chamberlin's generation of data that tested economic theory. I was front and center, acting in my own self-interest and from this construction arose deep insights into fundamental predictions from economic theory. Chamberlin's results puzzled me for years, keeping me up many nights, and it culminated in a life-long pursuit of using economic experiments in the laboratory, for which I was awarded the Nobel Prize in 2002, being recognized as the Father of Experimental Economics.

Prosecutor Naiv Ete: Very good, and congratulations on the Nobel Prize. Can you now tell me about your famous market study published in the *JPE* in 1962 for which the Nobel was awarded?

Vernon Smith: Of course. The Chamberlin design had my mind churning, and I honed in on the effect of institutions. To date, very little work was done exploring the effect of institutions, but my instincts suggested that the particular bilateral trading scheme of Chamberlin did not give neoclassical theory its best chance to succeed. I therefore wanted to engage in an efficacy test and give theory its "best shot". I decided to use the "double oral" auction, which had a group of buyers orally making bids, simultaneous to, and in clear sight, of sellers orally making asks. There was a centralized auctioneer that publicly displayed all bids and asks and cleared the market when there was a match. Beyond adding centrally occurring open outcry of bids and asks, I used multiple market periods, allowing students to learn during the auction sessions because I believed that learning in this new environment was important to reach equilibrium.

Empirical results from my double-oral auctions were staggering—quantity and price levels were very near competitive levels—and served to present the first evidence that Walrasian tâtonnement, conducted by a central auctioneer, was not necessary for market outcomes to approach neoclassical expectations. It is fair to say that this general result remains one of the most robust findings in experimental economics.

Prosecutor Naiv Ete: Can you describe why you believe these results are pathbreaking?

Vernon Smith: The bread and butter of economics is predictions from the supply and demand model. One pillar of economics is using empiricism to test our theories. But, the foundational information necessary—understanding both supply and demand and having the ability to systematically manipulate both curves—is very difficult to markets. Indeed, there are entire literatures in economics and marketing attempting to estimate demand curves with varying success. My laboratory experiments provide the necessary demand and supply information because I am inducing demand and supply curves for each subject, so I know them precisely. In addition, I can examine how the structure of the institution of trade affects the behavior of subjects as well as equilibrium outcomes.

Prosecutor Naiv Ete: And may I ask….were these student subjects that you used?

Vernon Smith: Yes, they were all Purdue undergraduate students who signed up to take part in a market experiment.

Prosecutor Naiv Ete: *Aha, uneasy lies the head that wears the crown Professor Smith.* Purdue? Signed up? So, these students might not even be representative of the broader Purdue undergraduate student population much less represent traders in the extra lab world? And, this market seems a completely abstract fictitious institution that you made up—would these students ever make such decisions in their normal course of life with such trading institutions?

Vernon Smith: What you say is mostly wrong—the market design was similar to the New York Stock Exchange, but more importantly you are missing the point of the study, which was to be an efficacy test of a foundational theory, and with my maintained assumption of parallelism I can be confident that…..

Prosecutor Naiv Ete: Enough Professor Smith, with all due respect, your "confidence" is illogical; this strikes me as scientific numerology in its finest form. You do not have representative people, situations, or stimuli—you're placing people on a decision margin that is entirely foreign

to them; so parallelism cannot hold, that is my point, and it is the correct one—in my opinion you have zero external validity. Trivial tautologies Professor Smith. I am done here your honor.

Judge Ima Learner: Defense, would you like to cross this witness?

Defender Minerva: Professor Smith, the Defense appreciates your unveiling of the empirical mystery behind supply and demand curves. In my first economics class at the University of Chicago with Professor Becker I had some trouble imagining the empirical content behind the curves because it was hard enough to estimate a single demand curve at one point in time, much less know the system and examine an equilibrium. But after reading your work it all became clear how one can reveal market behavior and test important neoclassical theories. Future generations of students can now appreciate the theoretical underpinnings of these curves and what they mean for markets. Generations of experimentalists have built on your broad shoulders. No questions from the Defense, Judge Learner.

Judge Ima Learner: Professor Smith, you are dismissed. The court thanks you for your time. Please call your next witness Mr. Ete.

Prosecutor Naiv Ete: The State calls Professor John List from the University of Chicago.

John List: I solemnly and sincerely declare and affirm that the evidence I shall give will be the truth, the whole truth, and nothing but the truth.

Prosecutor Naiv Ete: Thank you Professor List. Can you tell me about your experimental origins?

John List: Sure, back in the late 1980s I began to frequent sportscard trading markets to buy, sell, and trade baseball cards to earn extra spending cash for college. Throughout these days, I was constantly tinkering, combining my imagination with the logic of experimentation in every market we visited—it was always my first instinct to experiment. Given that such markets provided appealing naturally-occurring situations, in the early 1990s I began to run field experiments to

advance science. This early work attempted to figure out if the standard economic model could help us understand what was happening in these markets and to explore tools of non-market valuation, bargaining, and auctions. I learned early on that others did not appreciate the approach, with one anonymous journal referee noting "This author clearly does not understand experimental economics, which is best done in the laboratory…….I strongly advise rejecting this "field experimental" paper." Eventually this early work found a home, as did later 1990s sportscard field experiments (e.g., List and Shogren, 1998; List et al., 1998; List and Lucking-Reiley, 2000; List, 2001).

In reality, I never imagined sportscard markets being intrinsically interesting per se, rather they provided unique naturally-occurring settings where I could overlay randomization to answer interesting questions without putting people on an artificial margin. In a practical sense, it was my only venue for carrying out field experiments since I had no other means to fund my research. What I was essentially doing was using my baseball card collection—which I amassed with my paper route, snow shoveling, and grass cutting earnings from the 1970s and 1980s—to fund my academic research. In the past few decades, the budget constraint has loosened a bit and I have subsequently moved beyond sportscard markets and have used the world as my laboratory, publishing on issues as far removed as discrimination, market equilibria, gift exchange in the workplace, charitable giving, gender, energy conservation, the education production function, corporate social responsibility and the like; in many cases leveraging behavioral economics.

Prosecutor Naiv Ete: Very good. Behavioral economics seems an odd concept—I thought that was what economists did was examine individual behavior, but anyway can you now tell me about one of your early sportscard field experiments that explored an important aspect of behavioral economics?

John List: The behavioral landscape in the late 1990s was that preferences were not as maintained by standard models. The laboratory work in Kahneman et al. (1990) struck a deep empirical chord, reporting evidence of an endowment effect (or value disparity), which meant that people offer to sell a good in their possession at a substantially higher level than they will pay for the identical good not in their possession. Of course, claiming that preferences are defined over consumption levels or changes in consumption has serious implications for the discipline of economics. In a

normative sense, the basic independence assumption, which is used in most theoretical and applied economic models to assess the operation of markets, is directly refuted. In a positive sense, the disparity call into question commonly held interpretations of indifference curves, make cost-benefit analysis illegitimate, change the procedure necessary to resolve damage disputes, and cause a reconsideration of several of our other deep intuitions, such as the invariance result of Coase which relies on Hicksian equivalent surplus and Hicksian compensating surplus to be roughly equivalent.

From my viewpoint, this was all intriguing, but the most important questions facing such behavioral insights were whether they manifested themselves outside of the laboratory or likewise, whether our behavioral theories were adequate, and I…

Prosecutor Naiv Ete: Wait….so, you mean you want to explore whether the received laboratory results were externally valid Professor List?

John List: That is a start, but the issue is much deeper. My study (List, 2004) revolves around i) how agents actually trade and select into and out of naturally-occurring markets, ii) how market experience affects observed behaviors, and iii) if behavioral preferences are evident in markets, how does aggregation of those preferences alongside neoclassical preferences affect market outcomes? I find that, at odds with the literature's interpretation, the endowment effect is *not* a fundamental characteristic of preferences, as selection into markets and market experience, matter a great deal (as does aggregation). These market experience results have been broadly replicated in both the lab and the field, and there is now neurological evidence from fMRI supporting the results, providing a functional basis for the experience insights (Tong et al., 2016; see List, 2020, for a more patient overview of the evolution of this work).

Prosecutor Naiv Ete: *Aha, all that glisters is not gold* ***Professor List.*** I don't wish to sound so obtuse, but why should we care about behavior in magnetic tubes and sportscard markets? Sure, you are having people do what they naturally do in those markets, but frankly, like in Professor Smith's work, these people seem W.E.I.R.D: or as Professor Henrich testified, you have Western Educated Industrialized Rich and Democratic subjects (Henrich et al., 2010).

And, to make matters worse, these are data collected from convention centers, gymnasiums, hotels, conference headquarters, and the like where experimental control will be completely lost. I have frequented some of these places and they are more akin to madhouses than the crisp, clean, sterile environment of a well-controlled laboratory, or the egg-shaped enclosures that behaviorist psychologists used in the mid-1900s to provide "context free" temperature and sound regulated environments. Madhouses are not where data should be collected, Professor List, those areas are for madwomen and madmen who want to engage in behavior that is guided by no principles, a place where preferences are labile and subject to the whims of circumstance. Bespeak the egg, Professor.

John List: Well, the point of this line of work is not to learn about sportscard markets or the traders preferences therein per se, but to examine key economic theories in a natural setting where exogeneity permits a crisp look into…..

Prosecutor Naiv Ete: Enough Professor List, with all due respect, I don't want to waste more of your time….you have your precious baseball cards to tend to after all. I must admit that I find this whole baseball card field experimental approach point blank W.E.I.R.D—besides you, I have never met anyone who plays with baseball cards; I don't think we learn anything beyond sportscard markets. For that reason, I don't find any of this compelling, as you have zero external validity, and I see no real reason for conducting such drivel and publishing such balderdash. I am finished with this witness your honor.

Judge Ima Learner: Defense, would you like to cross this witness?

Defender Minerva: Professor List, the Defense thanks you for your creative use of markets across myriad of settings around the world—from classrooms to boardrooms to living rooms—to test economic theory and to learn about how economics can be used to enhance our understanding of human behavior. Indeed, upon taking an Uber to the courthouse this morning I realized that I was one of your experimental subjects when I considered which option to choose amongst the various tipping defaults. For the court's record, I am proudly in the Top 1% (see Chandar et al., 2019)! No questions from the Defense, Judge Learner.

Judge Ima Learner: Professor List, you are dismissed. The court thanks you for your time. Please call your next witness Mr. Ete.

Prosecutor Naiv Ete: The State calls Professor Esther Duflo, Professor of Economics at MIT.

Esther Duflo: I solemnly and sincerely declare and affirm that the evidence I shall give will be the truth, the whole truth, and nothing but the truth.

Prosecutor Naiv Ete: Thank you Professor Duflo. Can you tell me about your experimental origins?

Esther Duflo: When I first took a course in economics I didn't like it much. I was exposed to the typical Economics 101 course, the French version of it, with indifference curves, duality, and the entire mix of production and consumption ideas. I thought it had no relevance, that it was so far removed from a person's life that it could not possibly be useful. I was also skeptical that it was extremely ideological. While I have always had an analytic mind, I only took economics initially because I wanted to do economic history. So, I thought, "Okay, I'm done with this whole economics business." But at the same time, I was a little bit dissatisfied with history for different reasons, in particular, because I thought it was not very immediately useful. I was always of the mind that I wanted to change the world today not tomorrow.

So that's why I decided to go to Russia in 1993 under the pretext of writing my master's thesis, but really to get some exposure to the world. There, as a research assistant for Jeff Sachs, I got to see economists in action. I didn't necessarily agree with everything they were doing, but I could see that they had the luxury to think about things, and then to be heard by policymakers when they had something to say. And I said, "Wow, this is the best job in the world." And, about 10 years later after finishing my Ph.D. in Economics, I was myself having the luxury to think about how to address the most important questions using randomization in the developing world. In my early studies, I explored the role of information and social reactions in retirement plan decisions and how women performed as policymakers in India (Duflo and Saez, 2003; Chattopadhyay and Duflo, 2004).

In the latter natural experiment, I used unique natural randomization from a 1993 amendment to the constitution that required to reserve one-third of all positions of the chief to women. In this way, the local village councils (Gram Panchayats (GP)) were randomly selected to be reserved for women and therefore differences in investment decisions can be confidently attributed to the reserved status of those GPs. I found quite interesting results from this natural variation in the data. My work evolved into a life-long pursuit of making using of my own randomization in the developing world to tackle first order issue, for which I was awarded the Nobel Prize in 2019 (with Abhijit Banerjee and Michael Kremer).[5]

Prosecutor Naiv Ete: First of all, congratulations on your Nobel Prize. Second, can you tell me about your work for which the Nobel was awarded?

Esther Duflo: The Nobel Prize Committee awarded us the prize for using randomized control trials to fight global poverty. There are two steps to our approach. First, you take a very big, not very well-defined problem. For example, how can we end poverty? And then you try to break it into much more manageable questions, which can have a scientific answer. Like, what is the effect of tutors for low-performing pupils on test scores? This problem naturally pertains to the education production function, or how inputs map into outputs. If we are interested in increasing the output we measure as a test score, we need to understand what inputs are important.

Here is where the second step of this strategy comes in, which is to set up experiments to try to answer those questions. For example, in one of our experiments (Banerjee et al. (2007)) conducted in schools in urban India, we studied the effects of a remedial education program that hired young women to teach students lagging in basic literacy and numeracy skills. By randomizing the schools that received the program, we could examine how the remedial education program affected average test scores of all children in treatment schools.

Prosecutor Naiv Ete: Can you describe why you find this set of findings important?

[5] I heavily draw from https://www.npr.org/sections/money/2019/12/10/786431379/a-q-a-with-esther-duflo-who-wins-nobel-prize-today here and below.

Esther Duflo: I think that our findings are important because we identified a true barrier to student achievement: teaching methods that were insufficiently shaped to student need. Tutors for low-performing pupils in India improved achievement measurably, and lastingly. Importantly, this moves us one step closer to understanding what works in education and why. This is the first step to putting forward sound policies that can help children in developing countries. Also, this program was remarkably cheap and much more cost effective than hiring new teachers, which will make it more likely to be taken up by local governments. In the end, our results suggest that it may be possible to dramatically increase the quality of education in urban India.

Prosecutor Naiv Ete: And may I ask… were these interventions in various cities across India?

Esther Duflo: No, the experiment only took place in Vadodara with ninety-eight of Vadodara's 122 government primary schools.

Prosecutor Naiv Ete: So, students in these schools might not even be representative of the broader set of students in Vadodara much less represent the schools in rural India?

Esther Duflo: That is correct, but you are missing the point of the study, which was to…

Prosecutor Naiv Ete: *Aha, this is very midsummer madness Professor Duflo.* These program evaluation studies strike me as throwing a bunch of potential solutions at a problem and seeing if it works, in that moment for that setting. Such program evaluations might be socially good, but without theory testing or designs to explore whether the results will scale horizontally or vertically, this strikes me as program evaluation cherry picking at its finest. Find your favorite village to show efficacy and report beautiful results that mean nothing beyond your contrived sample. You do not have representative people, situations, or a means to scale these results to the level where it will matter. Indeed, I don't even see any mention of what aspects of the program could be scaled, which is an important piece of the external validity puzzle for policy research. Exhibit A for why policy science is an oxymoron. In my opinion you have zero external validity and zero chance to scale. You have learned a lot about the children of Vadodara, but what about children in Odisha, Yemen, Brisbane, Taipei, or Warsaw? And, what about if you had to double the subject pool in

Vadodara? Would the quality inputs be in place to secure the same treatment effects or would there be a voltage drop? We all know the answer. I am done here your honor.

Judge Ima Learner: Defense, would you like to cross this witness?

Defender Minerva: Professor Duflo, I applaud you for taking the field experimental method to the developing world. You are clearly changing this earth for the better, not only one village at a time ala the words of wisdom from Professor Curie, but by showing us all how we can embed science in the political process, which should be a primary public policy goal for decades to come. Systematically attacking problems step by step serves as a wonderful approach for learning, and to complement the other empirical exercises that take place in the social sciences. You make all of us, including Professor Curie, proud. No questions from the Defense, Judge Learner.

Judge Ima Learner: Professor Duflo, you are dismissed. The court thanks you for your time. Please call your next witness Mr. Ete.

Prosecutor Naiv Ete: The State calls Raj Chetty, Professor of Harvard University.

Raj Chetty: I solemnly and sincerely declare and affirm that the evidence I shall give will be the truth, the whole truth, and nothing but the truth.

Prosecutor Naiv Ete: Thank you Professor Chetty. Can you tell me about your empirical origins?

Raj Chetty: As I recall, I was first introduced to the empirical approach in economics thanks to Professor Martin Feldstein. Marty was such an inspiring figure for me in terms of the impact he has had on the profession, the questions he was working on, and how he was changing lives through his policy work. What struck me immediately was that the types of questions Marty asked and worked on were some of the most fundamental questions in economic policy. He didn't shy away from asking the big-picture questions, even though it might be difficult to answer them perfectly.

Later, in 2007, I came across an advertisement from the IRS seeking help organizing its electronic files into a format that would be easier to use for research. I immediately recognized that completing the job would make it possible for scholars to go far deeper into tax data. In the past, economists had to rely heavily on surveys, but the advent of cheap, powerful computing allowed for a new kind of economics—one that drew on the extensive administrative data gathered by governments. What began as a process of me and John Friedman registering to be federal contractors culminated with a life-long pursuit of using the empirical approach to solve the world's greatest problems with administrative data.

Prosecutor Naiv Ete: It is interesting that you questioned the wisdom that could be drawn from surveys, presumably believing that revealed preference was more valuable information than stated preference. I believe this is key to external validity concerns, and you seemed to realize that early in your career. So, let's go in that direction. Can you please tell me about your work for which you are most proud?

Raj Chetty: Wow, this is like picking your favorite child, but if I was forced to pick one, I would choose my work on the roots and consequences of economic and racial inequality, which spans a number of papers (see, e.g., Chetty et al., 2014, Chetty et al., 2018, Chetty et al., 2019.[6] The work broadly focuses on how we can give children from disadvantaged backgrounds a better chance of success. We use huge amounts of IRS tax data over a three-decade span to map inequality of opportunity in the US to the neighborhood level to show how American families fare across generations, revealing striking patterns of upward mobility and stagnation. My work shows that whether the American Dream is alive and well is indeed an empirical question.

[6] Much of these materials and information below are taken from:

https://www.vox.com/the-highlight/2019/5/14/18520783/harvard-economics-chetty

https://www.theatlantic.com/magazine/archive/2019/08/raj-chettys-american-dream/592804/

https://medium.com/conversations-with-tyler/raj-chetty-tyler-cowen-inequality-mobility-american-dream-d5ea7f4742b1

https://www.washingtonpost.com/business/economy/economic-mobility-hasnt-changed-in-a-half-century-in-america-economists-declare/2014/01/22/e845db4a-83a2-11e3-8099-9181471f7aaf_story.html

https://www.nytimes.com/2013/10/21/opinion/yes-economics-is-a-science.html

For example, the odds that a child reaches the top quantile of the national income distribution starting from a family in the bottom quintile is almost three times higher in San Jose compared to Charlotte. Analyzing the data we also find that African American boys in particular enjoy much less upward mobility than Caucasian boys, and that children born in 1940 had a 90 percent chance of earning more than their parents, but for children born four decades later, that chance had fallen to 50 percent, a toss of a coin. The American Dream really has changed, largely due to the economic opportunities available across the United States.

Prosecutor Naiv Ete: Can you describe why you find this set of findings important?

Raj Chetty: I find it important because the work presents a new portrait of intergenerational income mobility in the United States, one that was largely anecdotal before my work. When people talked about the "good old days" they refer to something that is actually real according to my data. There were the good old days in terms of an America that had many more opportunities for people to climb the ladder and realize the American Dream. In addition, we show that intergenerational mobility varies substantially across areas. Some regions of the United States look better than high-mobility countries such as Denmark, while others look more like a developing country. These stark differences recapitulate themselves on smaller and smaller scales as you zoom in. It is common to see opposite extremes of opportunity within easy walking distance of each other, even in two neighborhoods that long-term residents would consider quite similar. Our main lesson is that intergenerational mobility is a local problem, one that could potentially be tackled using place-based policies.

I also very much like the link that my more recent work makes back to my earlier research where I do a re-evaluation of the Tennessee STAR program. This is completed by combining IRS data with the experimental variation provided in the Tennessee STAR project, which was an experiment in the southern US (Chetty et al., 2011). This work shows the power of teachers within the education production function.

Prosecutor Naiv Ete: Very interesting. On the first line of work may I ask….were these data for all U.S. workers?

Raj Chetty: No, they were for people who filed tax returns from 1996-2012 in the US.

Prosecutor Naiv Ete: And, how about the Tennessee STAR project? What years and states did that data set cover and why is it important?

Raj Chetty: This was an experiment implemented across 79 schools in Tennessee from 1985 to 1989 and gives us an indication of the long-term impacts of early childhood education.

Prosecutor Naiv Ete: *Aha, I am a man more sinned against than sinning Professor Chetty*. So, these taxpayers might not even be representative of the broader U.S. workforce, much less the world? And, those data ended in 2012. My apologies Professor Chetty, but the US has changed fundamentally since then. Trumponomics, COVID-19, millions and millions of workers filing unemployment claims in 2020, technological advances making the GIG economy a key contributor to the workplace, and many more. Indeed, for the matter, the world is different, and you don't even address that. Now, as far as the Tennessee STAR results, these are educational programs from the 1980s; the family is different today, the labor market demands much different skills today, and the curriculum structures have fundamentally changed. And, we have not even gotten into the evidence that Tennessee STAR has failed dramatically when experts have tried to scale it. Indeed, it seems impossible to sample characteristics of the situation that have not occurred yet and are unpredictable—the type of information needed to scale such an intervention.

Raj Chetty: That is correct, but you are missing the point of the empirical exercise, which was to show how economic mobility and early human capital…

Prosecutor Naiv Ete: Enough Professor Chetty, with all the respect one can give another human, your time period was yesterday, not tomorrow, or even today. You have neither representative people nor situations—this means that you have no external validity. I wish you the best working on your spreadsheets and reams of IRS data that are uselessly invalid. You seem like a good chap, maybe you could scribe some useful theory or perhaps engage in a bit of successful thought experimentation? I am done Judge.

Judge Ima Learner: Defense, would you like to cross this witness?

Defender Minerva: Professor Chetty, the Defense thanks you for your wonderful insights across several areas of economics—from public economics to labor economics and nearly everywhere in between. Because of your work, we have a much better understanding of economic mobility, both its root causes and its underlying consequences. We now know how better to tackle such vexing issues and the world is better because of it. Keep up the great work Professor Chetty. No questions from the Defense, Judge Learner.

Judge Ima Learner: Professor Chetty, you are dismissed. The court thanks you for your time. Please call your next witness Mr. Ete.

Prosecutor Naiv Ete: The State calls our penultimate witness, Magne Mogstad, Professor from the University of Chicago.

Magne Mogstad: I solemnly and sincerely declare and affirm that the evidence I shall give will be the truth, the whole truth, and nothing but the truth.

Prosecutor Naiv Ete: Thank you Professor Mogstad. Can you tell me about your empirical origins?

Magne Mogstad: I was first introduced to the empirical approach in economics when I worked at Statistics Norway, which is the national statistical institute of Norway, and the main producer of official statistics. Statistics Norway collects data related to the economy and the society at national, regional and local levels and is one of the few statistical agencies in the world that has its own research department. When I worked there, we developed statistical methods in our analysis and I got interested in understanding each policy's unique impact so that we can provide credible empirical evidence that informs policy makers about which policies work and which ones do not. This culminated with a life-long pursuit of combining theory and econometric models with large

administrative datasets to answer the broad question of how to address market failures and equalize opportunities.[7]

Prosecutor Naiv Ete: Can you please tell me about your work for which you are the proudest?

Magne Mogstad: Top of mind for me right now is my most recent work with Marianne Bertrand and Jack Mountjoy on high school vocational education (Bertrand et al., 2019). Vocational education is one of the key avenues for augmenting human capital and there a lot of open questions. There is a perceived tradeoff between teaching readily deployable occupational skills versus shunting mostly disadvantaged students. We study the effects of a nationwide high school reform in Norway that aimed to move beyond this tradeoff. Reform 94 offered vocational students a pathway to college and sought to improve the quality of the vocational track.

We compared two groups of students—those born just after the January 1, 1978, cutoff date for eligibility and those born just before. We find that the reform increased initial enrollment in vocational-track high-school programs by more than 20 percent and even though enrollment in academic-track high-school programs decreased slightly, the reform nonetheless raised overall high-school matriculation. Overall, the reform reduced the gap in adult earnings between disadvantaged and less disadvantaged children by about 20 percent, and it was particularly effective at improving social mobility among men, with the gap in adult earnings between disadvantaged and less disadvantaged men decreasing by close to 30 percent.

Prosecutor Naiv Ete: Yes, we earlier heard from Professor Bertrand. She is quite impressive, so excellent choice of co-authors Professor Mogstad. Anyway, can you describe why you find this set of findings important?

Magne Mogstad: Our findings contribute to the debate on how secondary education can best be structured to improve social mobility. In the United States, vocational and technical education at the high-school level has long been controversial. Critics argue that vocational schools deprive

[7]Much of these materials and those below are drawn from https://review.chicagobooth.edu/economics/2019/article/how-norway-reduced-rich-poor-earnings-gap.

disadvantaged students of the opportunity to attend college and advocates maintain that vocational schools may better serve students who struggle with traditional academics, or those who can't, or don't, wish to attend college.

In recent years, however, a new vision has emerged, one that emphasizes increasing access to alternative educational models while ensuring that students who choose these pathways can still ultimately pursue higher education. In this sense, American reformers may find inspiration in our results.

Prosecutor Naiv Ete: And may I ask….were these data for all Norwegian citizens from the late 20th century?

Magne Mogstad: Yes, the starting point for our study is the Central Population Register, which contains records for every Norwegian from 1967 to 2015. We are very proud of the comprehensiveness of our data set.

Prosecutor Naiv Ete: This is worse than Smith, List, Duflo, and Chetty's garbage. First, Norway is what, the size of Mayberry, NC? Bailiff Fife, you are from that area of the world, am I right about its size?

Bailiff Barney Fife: Yes, I am from the lovely city of Mayberry, NC, and I had to apprehend lots of gangsters back in the day. It is my good sense that your estimate seems about right to me.

Prosecutor Naiv Ete: So, Professor Mogstad, who really cares about having every Norwegian in your *comprehensive* study? I could go tomorrow to Mayberry and have a comprehensive study. Second, your data ended in 2015. The world has changed a lot since 1967-2015. Just like Chetty, you seem not to understand that the world is much bigger than 2015 Mayberry, NC, Professor Mogstad. Bolsonaro, Trump, Boris, Macron, Greta Thunberg…lots has happened around the world Professor Mogstad. The world has changed.

Magne Mogstad: Jævel! You missed the point of the entire empirical exercise, which was to show…

Prosecutor Naiv Ete: *Aha, misery acquaints a man with strange bedfellows Professor Mogstad*. Professor Mogstad, with all due respect, let me ask you to leave now so you can go home and eat some herring with the other dozen or so Norwegian citizens you use for your *comprehensive science*. Your studies wouldn't pass an undergraduate course requirement for external validity much less something scientifically rigorous that journal editors or policymakers should consider seriously. I banish your work to the Journal of Last Resort (JLR) forever, or perhaps better yet the JNE, the Journal of Norwegian Economics. I bet a comprehensive group of 7 people will read it. I have nothing more to say your honor.

Judge Ima Learner: Defense, would you like to cross this witness?

Defender Minerva: Professor Mogstad, the Defense thanks you for both your empirical insights and your willingness to advance the very empirical techniques that we use to learn about the world. Your ability to tether empirical work with theoretical insights will continue to guide policymakers, theorists, and generations of scholars for years to come. Thank you, I wish I had a chance to learn from your work when I was a PhD student at the University of Chicago working with Professor Nerlove on random effects models; you would have undoubtedly enhanced our estimation approach and how to link it with important theories of the day. No questions from the Defense, Judge Learner.

Judge Ima Learner: Professor Mogstad, you are dismissed. The court thanks you for your time. Please call your next witness Mr. Ete.

Prosecutor Naiv Ete: The State calls our last witness, Professor Rebecca Diamond, Stanford University.

Rebecca Diamond: I solemnly and sincerely declare and affirm that the evidence I shall give will be the truth, the whole truth, and nothing but the truth.

Prosecutor Naiv Ete: Thank you Professor Diamond. Can you tell me about your empirical origins?

Rebecca Diamond: My father, Douglas Diamond, is a University of Chicago Professor, whose work focuses on banking and finance. Even with that, as a youngster I wasn't surrounded by wonderful economic theories and thinking through ways to use data to test those theories. Indeed, unlike some other children of academics, my father would not bring "home" his work and share it around the dinner table. In fact, I never really knew what he did until I went to college at Yale to study Physics, Economics, and Mathematics. I ended up earning a degree in Physics, Economics, and Mathematics at Yale, graduating *cum laude*. From there I went to Harvard to study for a PhD in Economics. This is where I was immersed in the empirical approach in Economics as a graduate student at Harvard, where I became interested in questions related to labor economics. Professors such as Lawrence Katz, Edward Glaeser, Ariel Pakes, and Claudia Goldin served as role models.

My main interests as a PhD student revolved around the implications of US workers' diverging location choices by skill level. In order to investigate the welfare implications of this divergence, I used data from the US census to obtain individual level observations, across three decades, on a wide range of economic and demographic information. Using this dataset including worker wages, housing costs, and geographic location of residence, I was able to estimate workers' choices.

Prosecutor Naiv Ete: This sounds fascinating. I trust that you were careful never to out-step the inference you made from your sample. Can you please tell me about your recent work in the gig economy with Uber?

Rebecca Diamond: Sure, that is work completed with Cody Cook and other colleagues (Cook et al., 2019). We used data from over a million Uber drivers, including their labor supply choices and earnings, to examine if we could find pay differences between men and women drivers. Surprisingly, we found that men who drive for Uber earn, on average, about 7% more than female drivers. Moreover, we were also able to show that this earnings gap between genders is explained entirely by differences in experience, choices over where and when to work, and preferences over

driving speed. We find no evidence that male and female drivers are differentially affected by a return to within-week work intensity or customer discrimination.

Prosecutor Naiv Ete: Can you describe why you find this set of findings important?

Rebecca Diamond: The findings in our paper are important because they suggest that there is no reason to suspect that the gig economy will help close gender pay differences, despite speculation that the increased worker flexibility will favor women. Since Uber pays drivers based on a non-negotiated formula that is invariant across drivers, our results show that even in the absence of discrimination and in flexible labor markets, women's relatively high opportunity cost of non-paid-work time and gender-based differences in preferences and constraints can sustain a gender pay gap.

In a related respect, a unique aspect of our data is our ability to both precisely measure a driver's experience and measure the return to experience through improved driver productivity, holding fixed the compensation schedule. Traditional datasets studying the gender pay gap often have very poor measures of experience (usually just a worker's age, sometimes years of employment). This measurement error in experience leads to attenuated estimates of the return to experience. We show that this measurement error in work experience can lead to biased estimates of the job-flexibility penalty. We know of no other work that shows this insight.

Relatedly, because drivers who work long hours also accumulate human capital at a faster rate per week, the importance of the job-flexibility penalty in the gender pay gap might be overstated in studies lacking good measures of worker experience. Separating out the importance of job-flexibility versus the return to experience for the gender pay gap in the broader economy is critical for formulating policy. Policies that improve job-flexibility (such as moving towards gig work) may only have a modest effect on the gender gap if the returns to on-the-job experience are a key driver of the hour-earnings relationship.

Prosecutor Naiv Ete: All very interesting. And may I ask….were these data from drivers in various cities across the US?

Rebecca Diamond: Yes, they were every UberX and UberPOOL drivers in the US.

Prosecutor Naiv Ete: So, you did not have other gig workers like freelance painters, accountants, tutors, or even Lyft drivers? And, your data ended in 2017?

Rebecca Diamond: That is correct, these are Uber drivers only from 2015 - 2017, but we think that this gives us a glimpse of the broader…

Prosecutor Naiv Ete: *Aha, lady doth protest too much, methinks*. Enough Professor Diamond. Clearly, you have not paid attention to these proceedings, or even read Campbell and Stanley's treatise on external validity. You do not have representative people, situations, or stimuli. We learn nothing from this particular empirical exercise since you have zero external validity. I doubt your findings would teach us anything about even GoPuff, Amazon.com, or Uber drivers, much less Google employees or the broader formal labor market of truckers, welders, and accountants. I will return to this issue in my closing statements Judge Learner but let's leave it here that this study is an exercise in well-wishing economists making a mountain out of a mole hill. I have made my relevant points your honor.

Judge Ima Learner: Defense, would you like to cross this witness?

Defender Minerva: Amazing use of "big data" to shed insights on the gender pay gap Professor Diamond. So far in your career I have watched with amazement as you tackle seemingly intractable problems with great wisdom and courage; a modern day Atalanta, slaying the deep problems of her day. I look forward to watching your career blossom even brighter because you are clearly a young superstar with a knack for tackling the most pressing issues of the day in an empirically-convincing manner. No questions from the Defense, Judge Learner.

Judge Ima Learner: Professor Diamond, you are dismissed. The court thanks you for your time. And, with that we can enter closing arguments. Please Mr. Ete, you can begin.

Prosecutor Naiv Ete: *To be, or not to be; that is the question Honorable Judge Learner*. Since this strikes me as an open and shut case, I will be brief your honor. Nearly 50 years ago Campbell and Stanley set forth arguments pertaining to external validity. They argued that several precepts must be met to satisfy such a critique. Over these past several years, I have ably taken the baton from these scholars and moved it to a higher plain. To a place that perhaps only the Illyrians have experienced.

As Professor Levitt made clear in his testimony, in 2007 he reminded us of generalizability when he and List questioned the value of the laboratory experiment to measure social preferences. Today, we are again reminded that all empirical exercises suffer from this grave generalizability bias that we call external validity.

Consider this last witness, Professor Diamond from Stanford Economics. We might call this encounter Exhibit A. She has Uber drivers across the US and claims that we learn something about the Gig economy and the broader labor market. I am left head scratching: are her results relevant for more traditional work contexts? Her work falls short of providing a convincing explanation of what can be extrapolated to more traditional labor market settings and, additionally, to contexts in which alternative work arrangements are prevalent.

We can ignore this kind of chicanery or we can eradicate it. I am a scientist, having trained in the outer reaches of Valhalla and UWSP, learning about the finer points of empirical work. I am here to exterminate it. It is the State of Disbelief's deep conclusions that the entirety of empirical economics suffers from external validity issues, and therefore every paper should suffer the fate of being published in journals rated no higher than the JLR, that is, the Journal of Last Resort. That is the fitting resting spot for these tools of weakness. In addition, empirical exercises should not guide the thoughts or decisions of any policymaker, organization, or firm, be it profit or non-profit.

Empirical work can never be universally trusted, so we must banish it from our society as we know it today. Your honor, it is time for you to make this planet great again. You can do so by taking us back to the pre-Baconian days, when life was much simpler and wisdom much deeper. Your honor, please help us make this planet great by exercising your authority and abating this noxious pollutant called empiricism. This pollutant must be curbed just like other dangerous toxics…arsenic, polonium, mercury and the entire botulinum family of neurotoxins.

To be, or not to be? The answer is simple your honor: *not to be*. The State rests.

Judge Ima Learner: Thank you Mr. Ete. Defense, your closing arguments please.

Defender Minerva: Thank you, Judge Learner. We have now been part of 1968 days of testimonies, having started back on September 25th, 2014. We have had brilliant witnesses visit this great court of Kanga Roo and detail their interesting scientific work. I was moved by both young and more mature scholars alike. Robert Metcalfe's work with the YMCA, Virgin Atlantic airline captains, Opower, and virtually any other firm that will have him is promisingly showing the scientific method is reaching every type of organization. Uri Gneezy's clever behavioral interventions teaches us how psychology complements economics to deepen our understanding of behavior in the wild. Professor Knight provided great wisdom when he reminded us that the existence of a problem in knowledge depends on the future being different from the past, while the possibility of a solution of the problem depends on the future being like the past (Knight 1921, p. 313). Except for the addition of a few new labels such as people and situations, his point represents acutely the crux of our external validity debate. The Prosecution would have you believe that in the face of such knowledge deficits, we should design incentives to do *fewer* empirical exercises. The Defense believes that this is the exact opposite of what a true learning society should do.

Prosecutor Ete's reasoning motivated me to pause, step back, and reconsider external validity in light of his closing arguments raising the work of Professor Diamond as his "Exhibit A". The Defense can imagine that reasonable people might object to the fact that these Uber results refer to very specific workers in a very specific occupational setting. Actually, I think that these features generate the most important value added that we can take out of this paper. Uber workers are completely free in deciding the timing, intensity, organization, physical capital and length of their work, with a complete job-related flexibility. Indeed, there is a literature (e.g., Goldin, 2014) opening up the possibility that if the workplace could be made more flexible, the gender wage gap might be reduced by allowing women to work with the flexibility that they need given their preferences and family commitments. Professor Diamond's study clearly shows that a perfectly flexible work environment is not enough to eliminate the gender gap in hourly compensation.

Let me emphasize once more the importance of this conclusion. It is hard (if not impossible) to find a better setting than the one studied in this paper to test convincingly the

relevance of the flexibility of the work environment for the gender gap. After this work, we can now focus more confidently on other factors explaining the gender gap, at least for relatively unskilled workers. The work also draws out a general point on the mis-measurement of work experience, and how that can have broader implications to past and future work.

Second, I want to reconsider the Prosecution's arguments on Levitt and List (2007a; 2007b), and clear up his misinterpretations, which seem to follow some others who have failed to read carefully the original study. Yes, on the one hand, Levitt and List (2007a) are very clear on what they are trying to accomplish in the study Mr. Naiv Ete cites (pp. 153-54):

> A critical assumption underlying the interpretation of data from many laboratory experiments is that the insights gained in the lab can be extrapolated to the world beyond, a principle we denote as generalizability. For physical laws and processes like gravity, photosynthesis, and mitosis, the evidence supports the idea that what happens in the lab is equally valid in the broader world. The American astronomer Harlow Shapley (1964, p. 43), for instance, noted that "as far as we can tell, the same physical laws prevail everywhere." In this manner, astronomers are able to infer the quantity of certain gases in the Sunflower galaxy, for example, from observations of signature wavelengths of light emitted from that galaxy.
>
> The basic strategy underlying laboratory experiments in the physical sciences and economics is similar, but the fact that humans are the object of study in the latter raises special questions about the ability to extrapolate experimental findings beyond the lab, questions that do not arise in the physical sciences. While few scientists would argue that observation influences whether Uranium239 would emit beta particles and turn into Neptunium, human behavior may be sensitive to a variety of factors that systematically vary between the lab and the outside world.

But, what Mr. Naiv Ete and some in the literature seem to ignore is that Levitt and List (2007a; 2007b) are quite optimistic that data used from lab and field experiments, and everywhere in between can yield important insights (as astute readers, such as Kessler and Vesterlund (2015) understood and endorse; pp. 170-71):

> The points we make concerning generalizability of lab data apply with equal force to generalizing from data generated from naturally occurring environments. Empirical economists understand that studies of sumo wrestlers or sports-card traders cannot be seamlessly extrapolated to other economic settings. Any empirical estimate requires an appropriate theory for proper inference—and this lesson holds whether the data are obtained in the lab, from coin collector shows, or from government surveys.
>
> ……one approach for the lab is to "nest" experiments one within another and then examine the different results of the related experiments. This approach may serve to "net out" laboratory effects and thus reveal more about deep structural parameters than running a simple, more traditional, experimental design. Additionally, lab experiments that focus on *qualitative* insights can provide a crucial first understanding and suggest underlying mechanisms that might be at work when certain data patterns are observed.

The Defense therefore holds that there is key value that can be gleaned from data gathered in unique settings, and while external validity concerns should be addressed, the burden of proof needs to rest with both authors and critics. Authors need to provide a playbook tackling upfront why external validity is not an important issue with their work, and critics must address fully why and how said external validity plays a detracting role. This needs to be not in an ad hominem fashion, but one that clearly states the principles used to determine said generalizability. We welcome guidance in this area Judge Learner. The Defense rests your honor.

Judge Ima Learner: Thank you Ms. Minerva. The court will take a 5-hour break and adjourn at 1700 local time for my final verdict.

……..

Bailiff Barney Fife: Order, order, all rise for the honorable Judge Learner.

Judge Ima Learner: Thank you Bailiff Fife. Over these past several years I have learned the details of the external validity argument and I have learned from the deep wisdom of experts, from the pioneers of knowledge creation to the new breed of scientists. What has been striking is that each expert considered the situation from the current empirical perspective of their day and asked about generalization and external validity. In this manner, a key illuminating feature that has been brought forward is that we never look at the world without the conceptual framework that has been built over time. Previous empirical observations, inductions, and introspections have importantly constructed that framework. The Bacon, Pasteur, and Fisher work built fundamental pieces of that framework that motivated a new conceptual framework used by later scientists. More parochially, within economics, the List work on market experience confirms some of the ways in which we look at the world. In the end, science strives to build knowledge. Science should never support generalizations of some incidental empirical observations. Inductions from empirical observations must prove themselves over time in endless tests and variations.

We now stand in a moment of significant experimental expansion within economics. Even so, through my discussions with my favorite co-editor of my favorite economics journal, the *JPE*, I have learned that external validity permeates roughly 90% of the more than several hundred empirical studies that he has handled in the past year. Now, what also seems clear is that

experimenters' work is especially susceptible to this issue because they have identification achieved via randomization—with naturally-occurring data, identification tends to be the showdown, yet of course, external validity is still oft-mentioned as we observed with Professors Chetty, Mogstad, and Diamond's work. What I have learned through these proceedings is that if you really desire, every empirical study can be rejected based using this complaint—time, situation, or population will cause you to fail every time, as Prosecutor Naiv Ete has argued for the past few years.

But I am a learner, and we should consider the scientific process as just that, a process to learn. Everything will not be exact, but each empirical exercise is a learning moment, and as we conduct new experiments, we should refine theories to organize the exact boundary conditions of our theory. To expect the original scholar to have every detail well understood is akin to expecting Heinrich Hertz's to know that laboratory size would have an important effect on observed radio waves, but that the other background factors would have no effect. It would defeat the advance of science to expect that each individual paper takes on every boundary condition; Rome is not to be expected to be built in one day. This issue demands structure, and I find that Defender Minerva is beginning to bring that structure through her wisdom. I will now lay out my own structure.

Theory Testing and External Validity

Received wisdom within the broader community is that tests of theory should be immune to external validity concerns because such studies are testing universal truths. Perhaps elucidating this view most clearly is Cook and Campbell (1979), who assert:

> The priority among validity types varies with the kind of research being conducted. For persons interested in theory testing it is almost as important to show that the variables involved in the research are constructs A and B (construct validity) as it is to show that the relationship is causal and goes from one variable to the other (internal validity). Few theories specify crucial target settings, populations, or times to or across which generalization is desired. Consequently, external validity is of relatively little importance.

This reasoning accords well with the argument that theory should be universal, so tests of it can occur in any given setting with any given population—if the theory is correct, it should manifest itself everywhere goes the hypothesis. While at first blush this line of argument seems reasonable, ultimately it is flawed for some types of theory testing.

Consider theory testing where the counterfactual, or baseline, is produced by the theory itself. That is, work where theory produces a "levels" prediction. In experimental economics, this would include common tests of allocation and bargaining games (e.g., dictator game, ultimatum game), sequential prisoner's dilemma games (e.g., trust game, gift exchange game), many market games (e.g., double-oral auction, bilateral bargaining, the suite of one-sided auction tests), social dilemma games (e.g., public goods and the many variants), and many tests of preferences (e.g., the endowment effect). In many sister sciences and the hard sciences, the numbers are much greater.

To gain a sense of my concerns with the Cook and Campbell (1979) assertion, consider tests of the endowment effect due to Kahneman et al. (1990; KKT hereafter) and List (2003). KKT (p. 1346) report that their data rejects neoclassical theory in that their "findings support an alternative view of endowment effects and loss aversion as fundamental characteristics of preferences." List (2003; 2004) reports results in concert with KKT when he considers inexperienced consumers. Yet, when he considers experienced agents, behavior matches neoclassical expectations, and therefore is at odds with the KKT conclusions.

So, KKT rejects neoclassical theory and advances the endowment effect theory. List (2003; 2004) finds some support for endowment effect theory but can reject strong endowment effects for experienced consumers. And, importantly, List's data cannot reject neoclassical theory for those types. If one takes the view that theory is universal, where does this leave us?

Importantly, work in the lab and field has replicated List's experience result (see citations in List, 2020), and new insightful theory has been introduced to help explain such behavioral patterns (e.g., Koszegi and Rabin, 2006). Yet, the process did not stop there. Tong et al. (2016) have gone back to the lab and dug a bit deeper into the experience results of List to produce neurological evidence supporting differences in brain activity between experienced and inexperienced traders, providing a scientific basis for those insights. This is just but one example, but all of these learnings would have been missed had we drawn the Prosecution's hard line that the original KKT study was not externally valid because it used students in an artificial setting.

To further organize thoughts about why external validity should be considered in such theory testing, I find it instructive to consider the general treatment effects approach when measuring the treatment effect, τ, in a simple model of the form: $Y = X\beta_1 + \tau T + \eta$, where Y is the outcome, X is a vector of domain specific factors (consider these background characteristics that affect Y), T is a treatment indicator, and η is random noise. In the endowment studies, for example,

Y is the propensity to trade. Since there is no *T* because everyone is simply making the same decision, the model becomes $Y = X\beta_1 + \eta$. These studies together illustrate the important point of how the *X* variation causes two very different results, and this is because of subject specific characteristics—market experience of the participants. The theory without the *X's* is incomplete.

Similar reasoning follows the entire body of work wherein theory provides the counterfactual. Walking through the same exercise with dictator game giving, for example, one might argue the relevant front-line test from theory is of the form: H_o: $Y = 0$; and since everyone gives zero in theory, a necessary condition is that $\beta_1 = 0$, though I have never seen that tested and discussed formally. In the allocation games, the relevant variables in *X* extend beyond individual features and include scrutiny, stakes, norms, social distance of partners, and the like (see Levitt and List, 2007a; 2007b).[8] In this case, the "deep" structural parameters obtained from a preference measurement experiment like a dictator game are dictated by environmental and person-specific factors, $X\beta_1$. Thus, background characteristics are again of import when testing the theory: when they are turned to one setting the theory will hold and when they are turned to another it will not hold.

If one considers testing theory wherein the data themselves provide the counterfactual (think of any general case where the examination pertains to measurement of a comparative static), τ can be measured with more confidence because the $X\beta_1$ term is differenced out due to randomization. In this manner, qualitative theoretical conclusions do not have the same external validity concerns as testing when theory produces a "levels" prediction.[9] If one wanted to agree with the wisdom of Cook and Campbell (1979) above, one would be on firmer ground doing so with such theory testing studies.

Vertical Versus Horizontal Generalizing

[8] They note (Levitt and List, 2007a; pp. 153-54): "In particular, we argue, based on decades of research in psychology and recent findings in experimental economics, that behavior in the lab is influenced not just by monetary calculations, but also by at least five other factors: 1) the presence of moral and ethical considerations; 2) the nature *and* extent of scrutiny of one's actions by others; 3) the context in which the decision is embedded; 4) self-selection of the individuals making the decisions; and 5) the stakes of the game."

[9] Of course, as List (2006) discusses, if one added an interaction term, $\tau T X\beta_2$, the treatment effect itself is a function of the environment. Likewise, even in the absence of such interaction effects, making inference about the treatment effect is difficult because in such cases the *X*s might be held constant at the "wrong" levels (i.e., a five-unit treatment effect estimate might be interpreted much differently against a baseline of three ($X\beta_1 = 3$) than against a baseline of three hundred ($X\beta_1 = 300$)).

One issue that I had not appreciated before our hearings commenced is the multi-faceted nature of generalizing. Consider first horizontal generalizing. The first step of such generalizing can be found in the recent replication movement in the social sciences: can we replicate results in identical populations and situations? This of course is not generalizing in any manner, but simply ensuring that we have found a true effect rather than a false positive. But it does set us up to generalize to different populations and situations that are situated similarly horizontally.

Consider the recent work of Fryer et al. (2015), who find positive evidence on the efficacy of their pre-K educational intervention in Chicago Heights, IL with parents (of 3-5 year olds) who are below the poverty-line. If we desired horizontal generalization of their insights (or likewise horizontally scale) we would be most confident in interventions that understood the "secret sauce" of the program, emulated those features, and did it in an environment that the theory viewed as "closely exchangeable." For example, if the secret sauce is that the original study employed the 20 best teachers in Chicago Heights, then we should be sure to hire the 20 best teachers in Dayton, Ohio if we scale to Dayton. Likewise, if a key environmental component is that the community is close-knit, then we would like Dayton to be close-knit because we expect the results to generalize because of exchangeability.

To generalize vertically (or likewise vertically scale), we would be interested in similar factors, but now generalizing takes a much richer form (see Al-Ubaydli et al., 2017a,b; 2019). Consider the Fryer et al. (2015) example. If the policymaker wanted to scale the program up to the entire southside of Chicago, we now have a different scaling problem than when we horizontally scale. This is because if the teachers are the key input, it might be much more difficult to hire 2000 teachers in the local labor market. In this case, you might need to hire "down" the quality scale, which causes the initial results not to scale, or if you want to maintain quality you might need to pay a higher wage, which would cause the benefit cost calculations from the original study not to scale. Either way, the results do not scale well, and this is an important generalizability consideration, particularly for policymakers.

Both of these cases highlight that generalization will be more seamless if the individual-specific variables (e.g., age, race, gender, socioeconomic status) as well as the environmental factors (teachers, peers, infrastructure, norms, etc.) that were present in the original setting that are theorized as relevant to the success of the intervention are also present in the target setting. Such aspects can be more readily understood if from the beginning the original researcher, policymaker,

and funder have a transparency checklist that can be used to give generalizability its best chance. I elaborate on my checklist to promote generalizability transparency in the next section.

A Transparency Checklist

Over the past several years I have learned that the two sides of this debate are quite entrenched, with strong and opposing views. Mr. Naiv Ete's desire to end all empirical work due to external validity concerns is certainly untenable, as I view the dark ages right around the corner from that decision. Likewise, the question of generalizability is not beyond dispute, as a combination of theory, introspection, and transparent empirical work can serve as a useful guide to deepening our understanding of the different people, situations, stimuli, and time periods that a particular result generalizes.

What strikes me as interesting is that while economics and many of the other social sciences are currently undergoing an external validity revolution, other areas are on the other end of the pendulum. For example, the evidence-based health practice literature seems to have lost focus on external validity. The irony of this seems lost on many of those in this field who wonder why science has such difficulty achieving application and widespread adoption of evidence-based practice. Indeed, our experts from the National Institutes of Health, the Center for Disease Control, and the Institute of Medicine all lamented that the percentage of evidence-based findings that have been translated into practice is discouragingly small. As Rothwell (2005) noted: "There is concern among clinicians that external validity is often poor [...]. Yet researchers, funding agencies, ethics committees, the pharmaceutical industry, medical journals, and governmental regulators alike all neglect external validity, leaving clinicians to make judgments." After these proceedings, I now understand why—they lack the structure to discuss the optimal knowledge path from bench to bedside.

All of this leads me to a checklist that will aid in the transparency of whether and to what extent the received results will generalize. When considering internal validity, the literature typically focuses on issues such as participant selection, attrition, compliance, and identification strategy. Such strategies correspond to structural, treatment-related, and observational differences between the treatment and control groups. These are features that every academic study now reports. I envision an analogous set of criteria to address external validity that every study should report to promote transparency.

Economics provides a useful set of tools to help us think about external validity. In the economic framework, individual choices depend on preferences, beliefs, and constraints. Observed differences in choice can be rationalized by each of these three types of variables. That should be our starting point for determining whether, and to what extent, behavior will change when moving across settings. For example, constraints include many relevant features, such as financial (with my current income can I afford that item) and non-financial considerations (with the current set of norms should I engage in that action). If we observe generous tipping behavior in face to face transactions between riders and cab drivers, should we expect a similar level of generosity when tipping is done privately? The answer is no if social norm adherence depends on the observability of behavior, as found in Chandar et al. (2019). As such, in settings where choice follows a norm, changing observability is expected to change choice, ceteris paribus. Similar thought experiments can be done with changing beliefs and preferences across settings.

With that economic structure in mind, the following 4 transparency conditions represent the burden of proof that The Supreme International Court on Empirical Methods places on authors:

Author Onus Probandi

1. SELECTION: Report clearly how selection of subjects occurred in 2 stages. First, provide details on the representativeness of the studied group compared to the underlying population.[10] Second, provide details of whether the study group is representative of the target population in terms of relevant observables that might impact preferences, beliefs, or individual constraints. If sorting into and out of the target market matters, then discuss how that might impact generalizability.
2. ATTRITION: Document attrition and compliance rates of subjects. Document reasons for attrition and non-compliance. Are there motivational or incentive differences between subject and target groups to maintain compliance?
3. NATURALNESS: Naturalness of the choice task, setting, and timeframe should be discussed. Does treatment reflect natural variability in task, choice setting, and timeframe as the target setting? Are subjects placed on an artificial margin when

[10] For example, a comparison of treatment and control outcomes from people who have volunteered and/or consented for a study might be different than a comparison of treatment and control outcomes of people who have not volunteered/consented, for example.

making the choice or in the timeframe of choice? Generally, is the nature of the choice and outcome architecture exchangeable between research and target settings (similar norms, stakes, and effort/outcome ratios as well as familiarity of choice, individual versus group decision, etc.)?[11]

4. SCALING: For those programmatic studies speaking to policymakers, the scientist should provide negotiables and non-negotiables if the program is scaled.[12] In other words, before scaling understand the program effects across subsets of the general population and characteristics of the situation to understand who should receive the program and where/how it should be implemented and whether it passes a benefit/cost test at scale.

I view these 4 SANS conditions as necessary, as in, sans discussing these, this Judge says "no thank you." Of course, studies are neither created equally nor for the same purpose. In this sense, weights of importance that one should place on each of these four burdens of proof during evaluation should be determined by the stage of research. A research hierarchy is needed, and the following represent waves of research:

WAVE1, EFFICACY AND PROOF OF CONCEPT: The basic building blocks of knowledge begin with exploratory work investigating causality, or efficacy, focusing primarily on producing first tests of theory or establishing initial causality. In such cases, external validity serves as an 'extra credit' component when the counterfactual is estimated with data. For instance, to begin building the knowledge base in an area, early design choices should be made to obtain high internal validity, such as use of a homogeneous population and a setting where treatment and control subjects can be treated with uniformity. Homogeneity within a population is usually obtained by restrictive inclusion and exclusion criteria, and comparability between groups and settings can be ensured via randomization. In this manner, WAVE1 serves as an efficacy test, or giving theory its

[11] In many cases, the answers to these questions are different for lab and artefactual field experiments compared to framed and natural field experiments, highlighting the complementary nature of these data generating processes (see Harrison and List, 2004; Al-Ubaydli and List, 2013; Czibor et al., 2019).

[12] Negotiables are design choices that can change at scale without affecting key outcomes. Non-negotiables are features that if changed at scale will considerably impact outcomes (see Al-Ubaydli et al., 2017b).

best shot, as well as exploring causal mechanisms. Nevertheless, an honest assessment of boundary conditions and useful further tests should be included in WAVE1 studies.

WAVE2, UNDERLYING MECHANISMS, BOUNDARIES, & REPLICATIONS: Building on the foundational features from WAVE1, WAVE2 studies, while maintaining the fidelity of internal validity, dig deeper into mechanisms, broaden the exploration of boundary conditions, and replicate. This can usefully start by relaxing homogeneity of population and situations. In this spirit, evaluating generalizability is of great interest with treatments conducted among a heterogeneous population and under heterogeneous conditions, particularly with concomitant factors present. These could vary subject populations, temporal dimensions, scrutiny, stakes, institutional rules, or any other factors deemed important theoretically. For studies in which measurement plays a key role, such as in medical trials and some social science trials, WAVE2 might include multi-arm studies, in which some arms represent the respective preference for the test or the control intervention – being an open and not blinded intervention – while the other arms are the randomized, blinded trials.

The distinction between empirical designs in WAVE1 and WAVE2 might be viewed as the difference between science carried out in an "ideal" versus "everyday" (or "policy") conditions. That is, when switching the focus from the confirmation of an efficacy trial to the question of a broader application of an intervention, the design choices vary dramatically. In WAVE1, where often the goal is to give the theory or intervention its best chance, it is important to have ideal conditions such as well-trained and highly experienced teachers, administrators, therapists, etc., a population that is theoretically most moved by the intervention, and a setting that ensures optimal compliance and minimal attrition. In WAVE2, adding realistic factors that mirror natural settings, adding concomitant natural interferences, and other contraindications to the setting as an explicit manner to address generalizability are important. An ambitious original study might even combine WAVE1 and WAVE2 research during the WAVE1 phase.

WAVE3, MEASUREMENT, MECHANISMS, & SCALING: Should be viewed as the final research completed before policy implementation or a deep understanding of the magnitude of the treatment effect, the underpinnings for why the intervention works, and a description of important

boundary conditions. Key work for policymakers in WAVE3 focuses on what is necessary to scale the received results. Researchers should backward induct when setting up their WAVE3 research to ensure accurate and swift transference of programs to scale. The checklist should include information from the Author Onus Probandi, as well as a complete cataloguing of both benefits and costs, with estimates of how those will change at scale. To learn about these elements, researchers should block on situations when doing experiments, just like we commonly block on individual characteristics in modern experimentation. This might come in the form of leveraging multi-site trials to learn about the variation of program impacts across both populational and situational dimensions. WAVE3 holds the empirical promise of moving the area of study from evidence-based policy to policy-based evidence.

Final Verdict

Many critics view unique empirical settings as a negative distraction. This is the exact opposite way to think of the issue: if the uniqueness itself allows you to do the relevant test and no other setting can achieve that level of relevance, then you have found the "perfect" domain for your study, not an imperfect one.

I urge all writers to take themselves through an EV Litmus Test. Can you confidently say that it is quite difficult, if not impossible, to find a better setting than the one studied in your paper to test convincingly the relevance of your conjectures? If so, then state that fact. This is a key consideration. If not, then state the relevant potential substitute settings alongside the transparency evidence in your study. If your transparency evidence reveals that your empirical approach diverges from the target setting on certain dimensions of import, your theory should provide a framework for predicting in what direction behavior in your setting will deviate from that of the target setting.

All results are externally valid to some setting, and no result will be externally valid to all settings. Detailing similarities and differences between the study setting and target settings in terms of relevant factors that might impact preferences, beliefs, or individual constraints is a key consideration for both advocates and proponents of the empirical method.

Once the author satisfies the Onus Probandi, then the burden of proof is shifted to the detractor to identify factors that are indeed relevant to the shaping of the phenomenon that were observed in the research setting that are different from the target setting. The sources of external

invalidity must be much more than guesses as to general laws in the science of a science: insights on what factors forcefully interact with treatment, and, by implication, affect outcomes must come from theory not the heart. The detractor should rely on features associated with preference, belief, and constraint differences between the research and target settings. In this manner, in much the same way that economists need a model of firm and consumer behavior to inform us of the parameter we are estimating when we regress quantities on prices, we need a behavioral model to describe the data-generating process, and how it is related to other settings. Theory is the tool that permits us to take results from one environment to predict in another, and generalizability, and criticism thereof, should be no exception.

Through these proceedings I often found myself reflecting on Tolstoy's Anna Karenina, and the resulting named principle. Generalization is a fragile concept, as is the narrower concept of whether a program scales. Any one of a number of factors can doom such exercises, consequently, a successful generalization is one where every possible deficiency has been avoided. I believe that we can use the Anna Karenina Principle to understand why so few programs have successfully scaled through history. Under this line of argument, all successfully scaled projects are not so because of a particular trait, but because of a lack of any number of possible negative traits. In this manner, disparate preferences, beliefs, and/or constraints across the research and target settings are each expected to frustrate successful generalization (scaling).

For critics, such as Prosecutor Naiv Ete, who argue that when a shadow of an external validity doubt creeps in, journal editors and policymakers should extinguish the evidence because there are no learnings possible, I urge you to rethink your positions. Much at odds with Prosecutor Naiv Ete, I view the four tenets of empiricism as follows:

1. Theory and empiricism are symbiotic: theory provides a structure for thinking about the world, empirical work tests whether that structure is approximately correct and informs future theories.
2. One swallow does not make a summer: each study moves priors by an amount corresponding to its quality and the strength of priors.
3. To explain differences in observed choices across settings, ask if preferences, constraints, or beliefs have changed.
4. Uniqueness of a setting can be a key strength, not a weakness, if it isolates a particular channel or causal mechanism effectively.

Bailiff Barney Fife: The honorable Judge Learner of the Court of Kanga Roo has spoken. Ye empiricists shall go forth and generate data in unique settings that continue to illuminate the ways of humankind.

Prosecutor Naiv Ete: *I came, I saw, and overcame. Yet, this unkindness may defeat my life, good night, good night…parting is such sweet sorrow.*

References

Al-Ubaydli, O., & List, J. A. (2013). "On the Generalizability of Experimental Results in Economics." In Frechette, G. & Schotter, A., *Methods of Modern Experimental Economics*, Oxford University Press.

Al-Ubaydli, O., List, J. A., LoRe, D., & Suskind, D. (2017a). "Scaling for Economists: Lessons from the Non-Adherence Problem in the Medical Literature." *Journal of Economic Perspectives*, *31*(4), 125–144. https://doi.org/10.1257/jep.31.4.125

Al-Ubaydli, O., List, J. A., & Suskind, D. (2019). "The Science of Using Science: Towards an Understanding of the Threats to Scaling Experiments." *International Economic Review*, forthcoming, and NBER Working Paper No. 25848.

Al-Ubaydli, O., List, J. A., & Suskind, D. L. (2017b). "What Can We Learn from Experiments? Understanding the Threats to the Scalability of Experimental Results." *American Economic Review P&P*, *107*(5), 282–286.

Bazin, Hervé. "Pasteur and the birth of vaccines made in the laboratory." In *History of Vaccine Development*, pp. 33-45. Springer, New York, NY, 2011.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden. "Remedying education: Evidence from two randomized experiments in India." *The Quarterly Journal of Economics* 122, no. 3 (2007): 1235-1264.

Bertrand, Marianne, Magne Mogstad, and Jack Mountjoy. *Improving educational pathways to social mobility: Evidence from Norway's "Reform 94"*. No. w25679. National Bureau of Economic Research, 2019.

Box, Joan Fisher. "R. A. Fisher and the Design of Experiments, 1922-1926." *The American Statistician* 34, no. 1 (1980): 1-7.

Brunswik, E. "The conceptual framework of psychology." In *International encyclopedia of unified science* (Vol. 1, no. 10, pp. 4-102). (1952) Chicago: University of Chicago Press.

Brunswik, E. *Perception and the representative design of psychological experiments*. (1956; 2nd ed.). Berkeley: University of California Press.

Campbell, Donald. T., and Julian C. Stanley. Experimental and quasi-experimental designs for research. Chicago: Rand McNally & Company, (1963).

Chamberlin, Edward H. "An experimental imperfect market." *Journal of Political Economy* 56, no. 2 (1948): 95-108.

Chandar, Bharat, Uri Gneezy, John A. List, and Ian Muir. "The Drivers of Social Preferences: Evidence from a Nationwide Tipping Field Experiment" *Field Experiments Website, working paper.* (2019).

Chattopadhyay, Raghabendra and Esther Duflo. "Women as Policy Makers: Evidence from a India-Wide Randomized Policy Experiment." *Econometrica* 72, no. 5 (2004): 1409- 444.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." *Quarterly Journal of Economics* 126(4): 1593-1660 (2011).

Chetty, Raj, John N. Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan. "Income segregation and intergenerational mobility across colleges in the united states." *The Quarterly Journal of Economics* (2019).

Chetty, Raj, John N. Friedman, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter. *The opportunity atlas: Mapping the childhood roots of social mobility*. No. w25147. National Bureau of Economic Research, 2018.

Chetty, Raj, David Grusky, Maximilian Hell, Nathaniel Hendren, Robert Manduca, and Jimmy Narang. "The fading American dream: Trends in absolute income mobility since 1940." *Science* 356, no. 6336 (2017): 398-406.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. "Where is the land of opportunity? The geography of intergenerational mobility in the United States." *The Quarterly Journal of Economics* 129, no. 4 (2014): 1553-1623.

Chetty, Raj. "Yes, economics is a science." Opinion. The New York Times. October 20, 2013. https://www.nytimes.com/2013/10/21/opinion/yes-economics-is-a-science.html

Cook, Cody, Rebecca Diamond, Jonathan Hall, John A. List, and Paul Oyer. *The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers*. No. w24732. National Bureau of Economic Research, 2018.

Cook, Gareth. "The Economist Who Would Fix the American Dream." The Atlantic. July 17, 2019.

Cook, Thomas D., and Donald T. Campbell. "The design and conduct of true experiments and quasi-experiments in field settings." In *Reproduced in part in Research in Organizations: Issues and Controversies*. Goodyear Publishing Company, 1979.

Cowen, Tyler. "Raj Chetty on Teachers, Taxes, Mobility, and How to Answer Big Questions (Ep. 23)." Conversations with Tyler. Medium. July 17, 2019. https://medium.com/conversations-with-tyler/raj-chetty-tyler-cowen-inequality-mobility-american-dream-d5ea7f4742b1.

Czibor, Eszter, David Jimenez-Gomez, & John A. List, "The Dozen Things Experimental Economists Should Do (More Of)," *Southern Economic Journal*, 86 (2), (2019): 371-432.

Dixon, Bernard. “The hundred years of Louis Pasteur”. *New Scientist*. No. 1221. Reed Business Information, 30–32 (1980).

Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. "Using prediction markets to estimate the reproducibility of scientific research." *Proceedings of the National Academy of Sciences* 112, no. 50 (2015): 15343-15347.

Duflo, Esther, and Emmanuel Saez. "The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment." *The Quarterly journal of economics* 118, no. 3 (2003): 815-842.

Fisher, Ronald A., and Winifred A. Mackenzie. "Studies in crop variation. II. The manurial response of different potato varieties." *Journal of Agricultural Science* 13, no. 3 (1923): 311-320

Fryer Jr, Roland G., Steven D. Levitt, and John A. List. *Parental incentives and early childhood achievement: a field experiment in Chicago heights*. No. w21477. National Bureau of Economic Research, 2015.

Gradmann, Christoph. "A harmony of illusions: clinical and experimental testing of Robert Koch’s tuberculin 1890–1900." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 35, no. 3 (2004): 465-481.

Gunn, Dwyer. “How Norway Reduced the Rich-Poor Earnings Gap.” Chicago Booth Review. September 24, 2019. https://review.chicagobooth.edu/economics/2019/article/how-norway-reduced-rich-poor-earnings-gap.

Harrison, Glenn, W., and John A. List. “Field Experiments.” *Journal of Economic Literature* 42, no. 4 (2004): 1009-1055.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. “The weirdest people in the world?” *Behavioral and Brain Sciences* 33, no. 2-3 (2010): 61-83. doi:10.1017/s0140525x0999152x

Hill, Austin Bradford. “The environment and disease: Association or causation?” *JRSM* 58, no. 5 (1965).

Hotz, Joseph V., Guido Imbens, and Julie Mortimer. “Predicting the efficacy of future training programs using past experiences at other locations.” *Journal of Econometrics* 125, no. 1-2 (2005): 241-270.

Ioannidis, John P. “Why Most Published Research Findings Are False.” *PLoS Medicine* 2, no. 8 (2005).

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. "Experimental Tests of the Endowment Effect and the Coase Theorem." *Journal of Political Economy* 98, no. 6 (1990): 1325-348.

Kessler, Judd, and Lise Vesterlund. *The external validity of laboratory experiments: The misleading emphasis on quantitative effects*. Vol. 18. Oxford, UK: Oxford University Press, 2015.

Kőszegi, Botond, and Matthew Rabin. "A model of reference-dependent preferences." *The Quarterly Journal of Economics* 121, no. 4 (2006): 1133-1165.

Knight, Frank H. "Cost of production and price over long and short periods." *Journal of Political Economy* 29, no. 4 (1921): 304-335.

Goldin, Claudia. "A grand gender convergence: Its last chapter." *American Economic Review* 104, no. 4 (2014): 1091-1119.

Levitt, Steven D., and John A. List. "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives* 21, no. 2 (2007a): 153-174.

Levitt, Steven D. and John A. List, "Viewpoint: On the Generalizability of Lab Behaviour to the Field." *Canadian Journal of Economics*, 40, No. 2 (2007b), pp. 347-370.

List, John A. and Jason Shogren. "Calibration of the difference between actual and hypothetical valuations in a field experiment." *Journal of Economic Behavior and Organization* 37, no. 2 (1998), 193-205.

List, John A. and David Lucking-Reiley. "Demand Reduction in Multiunit Auctions: Evidence from a Sportscard Field Experiment." *American Economic Review* 90, no. 4 (2000): 961-972. doi:10.1257/aer.90.4.961

List, John A. "Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sportscards." *American economic review* 91, no. 5 (2001): 1498-1507.

List, John A., "Does Market Experience Eliminate Market Anomalies?" *Quarterly Journal of Economics*, (2003), 118(1), pp. 41-71.

List, John A. "Neoclassical theory versus prospect theory: Evidence from the marketplace." *Econometrica* 72, no. 2 (2004): 615-625.

List, John A. "Field Experiments: A Bridge between Lab and Naturally Occurring Data," *The B.E. Journal of Economic Analysis & Policy*, (2006), 6(2 - Advances), Article 8.

Maniadis, Zacharias, Fabio Tufano, and John A. List. "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects." *American Economic Review* 104, no. 1 (2014): 277-90.

Matthews, Dylan. "The Radical Plan to Change How Harvard Teaches Economics." The Highlight by Vox. Vox. May 22, 2019.

Rothwell, Peter M. "External validity of randomised controlled trials:“to whom do the results of this trial apply?”." *The Lancet* 365, no. 9453 (2005): 82-93.

Tankersley, Jim. "Economic mobility hasn’t changed in a half-century in America, economists declare." Business. The Washington Post. January 23, 2014.

Tong, L.C., Karen, Ye, Kataro Asai, Seda Ertac, J.A. List, Howard Nusbaum, and Ali Hortaçsu, 2016. Trading experience modulates anterior insula to reduce the endowment effect. *Proceedings of the National Academy of Sciences*, *113*(33), pp.9238-9243.