This is a peer-reviewed, author's accepted manuscript of the following research article: Ting, C., & Montgomery, M. (2023). Taming human subjects: researchers' strategies for coping with vagaries in social science experiments. *Social Epistemology*, 1-17. https://doi.org/10.1080/02691728.2023.2177128

Title

Taming Human Subjects: Researchers' Strategies for Coping with Vagaries in Behavioral Experiments

Author details:

Carol Ting (corresponding author)
ORCID: 0000-0002-1135-5689
Department of Communication, Faculty of Social Sciences, University of Macau

Martin Montgomery
ORCID: 0000-0002-8426-4817
1. School of Humanities, University of Strathclyde
2. Department of English, University of Macau

Taming Human Subjects: Researchers' Strategies for Coping with Vagaries in Social Science Experiments

Abstract

The experimental method is designed to secure the reliable attribution of causal relationships by means of controlled comparison across conditions. Doing so, however, depends upon the reduction of uncertainties and inconsistencies in the process of comparison; and this poses particularly significant challenges for the behavioral and social sciences because they work with human subjects, whose malleability and complexity often interact in unexpected ways with experimental manipulations, thus resulting in unpredictable behavior. Drawing on the Science and Technology Studies perspective and one of our author's experiences in experimental work, this paper examines how experimental social scientists manage to establish objectivity and standardization in the face of vagaries arising from working with human subjects. In identifying experimental researchers' solutions to this challenge, we draw on methodological discussions among applied social scientists as naturally occurring data, through which we show how some seemingly mundane practices play essential roles in extracting patterns out of otherwise unpredictable behaviors in the lab. Closely examining such strategies, we reveal the inherent instabilities in the experimental method when adopted in the social sciences and discuss their methodological implications. In conclusion, we make tentative suggestions for escaping the kinds of methodological impasse which we have identified.

Keywords: experimental method, auxiliary assumptions, human subjects, replication crisis

Introduction

The capacity to enable causal inference is generally taken to be the strong suit of the experimental method (Carr et al. 2018; Guala 2005; Pearl and Mackenzie 2018). Compared to researchers working with other methods, experimenters have more control over their observational conditions: by inducing phenomena in a lab, experimenters can exclude or limit certain types of noise, reproduce the phenomena, and tinker with them. This ability to tinker with phenomena in the lab helps researchers learn about the phenomena (Hacking 1983; Pickering 1995), which is the epistemological foundation of experimental sciences.

However, the extent to which it is possible to tinker with phenomena in the lab is not uniform across disciplines. Chemists and physicists can experiment with inanimate/insensate materials without worrying about changing the nature of the underlying phenomena because physical entities and chemicals have stable properties and act with relatively predictable patterns. On the other hand, experimental behavioral and social scientists[1] work with human subjects whose nature is fundamentally different. Humans are highly heterogeneous and malleable; they come with idiosyncrasies and they often interact with experimental manipulations in surprising ways. Experimental subjects may be sitting alone in a quiet cubicle without contact with the outside world, but their behavior could be influenced by everything big and small that they had experienced in the past day, the past week, the past year, and so on. They are aware of being in an experiment and they may react to attempts to manipulate their behavior and attitudes in various ways (Gibson 2019; Marowski 2015;

[1] By "experimental behavioral and social sciences" we mainly refer to disciplines that study human behavior with the experimental method. According to conventional disciplinary division, most experimental social sciences falls into this category, but some areas not conventionally considered as social sciences, such as neuroscience, also fit this description. In the rest of the paper we use the general label "social sciences" for succinctness.

Rosenthal and Rosnow 2009; Stam, Radtke, and Lubek 2000). Since behavioral and social sciences are modeled after natural sciences (more accurately, physics, the most prestigious of all),[2] this difference in subject nature creates tension: how do experimental researchers of human behavior meet the expectations of objectivity and standardization with lively and often unpredictable human subjects? Surprisingly, to date this area of research has received rather limited attention, and it is this interesting question that we want to highlight in this paper.

There is a small but growing body of research on the challenges of conducting social science experiments. For example, based on ethnography and interviews, Cohn (2008) describes the difficult balance neuroscientists must strike by directing subjects to follow laboratory scripts on the one hand while keeping the appearance of measuring mental processes isolated in the brains of subjects on the other. Also taking an ethnographic approach, Peterson (2015) compares practices in psychology and molecular biology labs and explains the lack of progress in experimental psychology as stemming from the difficulty in creating and stabilizing new manipulation techniques and technologies that would help push the research frontier (see also Peterson 2016). Ting and Fitzgerald's case study (2020) on how pilot runs inform experimental designs provides many examples of 'troubles' with predicting subject behavior and the critical role of iterative design tweaks in stabilizing the pattern under study. Aiming to study the role of experimental tasks, Morrison and colleagues (2019) use a mind-wandering experiment to elicit the subject's experience through interviews. Accounts of the experience of subjects suggest a picture of complexity and

[2] We acknowledge the heterogeneity across scientific sub-disciplines and that not all sub-fields of natural sciences can reach the same level of mathematical sharpness and predictive precision of physics (Nelson 2016). In the rest of the paper we sometimes use the wording "natural sciences" or "the hard sciences" to refer to the older and more deterministic sciences including Newtonian physics and chemistry, but our comparison is a very targeted one that aims to focus attention on the challenges specific to experimental social sciences.

unpredictability that contradicts accurate measurement and systematic interpretation. Similar difficulties are also evident in Gibson's analysis (2019) of original recordings from Milgram's obedience experiment as he shows how subjects actually 'talked back' in various ways not acknowledged in the original report.

Two recent studies on experimental economics are particularly noteworthy because they directly look at the construction of order from an interactionist point of view. Böhme's ethnomethodological study (2016) draws from observational data of lab experiments and shows how, through normative expectations, monetary incentives and lab instructions, experimental economists make lab subjects perform the role of the rational maximizer which is the nexus between economic theory and experiments. More recently, Asdal and Cointe (2022) focus on written lab instructions as text-devices that move through all stages of the publication process in experimental economics. Through interview data with experimental economists, Asdal and Cointe argue that written instructions are essentially material and semiotic resources that facilitate cooperation in the lab and collective validation in the discipline.

These studies use observational and interview data, which have been the most powerful tools for Science and Technology Studies (STS) researchers. However, these kinds of data have limitations: access to experimenters' practices through these methods are partial in the sense that they can reveal only practices that are visible in the lab and/or are acknowledged by interviewees. As a result, methodological controversies and difficulties that are deeply embedded in social science experiments sometimes escape the inquiry of observers and interviewers (Danziger, 1990, p. 13). For example, it has been shown in experimental economics that, when presented with two risky bets, subjects exhibit different preferences depending on the given response format (choice vs. naming prices), and a whole body of influential paradigmatic research is contingent on having subjects choose between

two options rather than negotiating prices (Slovic 1995; Tversky, Slovic, and Kahneman 1990). Similar patterns were also demonstrated in experimental psychology and in survey studies (Birenbaum and Tatsuoka 1987). This phenomenon, however, is little known outside the circle of methodologists; and such dependence of experimental outcomes on response formats remains underappreciated.

Response format dependence is just one example of practices in experimental social sciences that are typically invisible to both practitioners and outsiders. The mundane and seemingly peripheral appearance of these practices obscures their important role and keeps them from the view of those seeking to understand the experimental machinery of social sciences. One good place to find such practices, though, is in methodological discussions on experimental economics and psychology, where reflective practitioners often voice their misgivings and debate the connection between their methods and findings. Insights from these sources can complement those gleaned from observational and interview studies. Drawing on the first author's training as a quantitative researcher and her experience of working on behaviorial experiments, we therefore take the rather unconventional approach of collecting data from methodological literatures for this paper on social science experiments.

This approach yields two main insights. First, comparative and discipline-specific methodological discussions in economics and psychology reveal fundamental ways in which theoretical abstraction shapes experimental practices. Second, in addition to theoretical abstraction, experimental social scientists converge on some often-overlooked practices for accomplishing objectivity and standardization, which we designate under the following headings of *purifying lab subjects*, *restricting response space*, and *removing contaminated responses*. While these strategies underpin how the experimental method is implemented with human subjects, their ability to compensate for inherent instabilities resulting from heterogeneous and malleable human behavior remains limited. As a result, conclusions based

on experiments can end up being challenged by exposing their taken-for-granted, tacit working practices. And this, ultimately, makes the task of consensus-building in the social sciences problematic.

Implementing the Experimental Method in the Social Sciences

The experimental method is often touted as the gold standard of scientific research because of its strength in causal attribution. This strength, in turn, is contingent on controlled comparison (a.k.a. the *ceteris paribus* principle; Cartwright 1983): if two conditions differ *only in experimental manipulation*, the difference in outcomes can then be attributed to the manipulation. Although the idea of controlled comparison sounds compelling, in practice it is never guaranteed that the experimental conditions differ only in manipulation. If researchers had understood everything about a phenomenon, there would be no need to further experiment on it; but if researchers do not fully understand the phenomenon, chances are that some causally-relevant factors are unknown and uncontrolled for. Causal attribution is impossible without overcoming this circularity, which in turn requires a leap of faith: the researcher has to assume that key causally-relevant factors are known and controlled for – such assumptions are called *auxiliary assumptions* (Guala 2005). Only with auxiliary assumptions can researchers proceed to treat their experimental setup as a closed system and apply the principle of controlled comparison for causal attribution.

Of course, a leap of faith can be nothing but dangerous if it is based on wishful thinking. When researchers fail to control for causally-relevant factors, the difference in experimental outcomes will be contaminated by these factors (or *confounds*, as they are known[3]) and the power of causal attribution is thereby undermined. Therefore, experimenters

[3] Statisticians debate about how best to define confounds (Pearl and Mackenzie 2018). Some emphasize a spurious variable connected to both the input and output variables (e.g., Carr et al. 2018), but the word confound

dedicate much effort to the control of potential causally-relevant factors. Nevertheless, given that the experiment setup always relies on auxiliary assumptions, no experimental tests can conclusively prove a hypothesis because the outcome can always be explained away by some of the auxiliary assumptions not being met. This difficulty is the well-known Duhem-Quine problem (Sismondo 2010). When applied to the context of scientific controversies, it manifests itself as the "experimenters' regress" as Collins (1985) coined it: the differences in experimental outcomes can always be attributed to some (often subtle) differences in experimental implementation. Since the causal relevance of such implementational differences often cannot be assumed *a priori*, it follows that no single experiment can be the ultimate arbiter of truth. In other words, the 'facts' established with the experimental method always contain interpretations on top of data (Teo, 2010).

Confounds, or unaccounted-for causally-relevant variables, pose a significant challenge to social science experiments. Being highly malleable and heterogeneous, humans make for difficult experimental subjects compared to inanimate matters like physical objects or chemical elements, which behave much more predictably – a hydrogen atom (or any other fundamental particles) acts the same way today as 100 years ago or later, but human behavior can change rapidly in complex ways. Moreover, human subjects bring into the lab their cultural norms, history, personality, mood and cognitive style, which are often extraneous to the researchers' theoretical model. For example, in an economics paper on cooperative behavior, the author explains an anomaly in the data in this way:

> Session 10 has high contribution levels in both treatments relative to other sessions with the same matching protocol. At least a partial explanation for this is that on the day of the experiment some subjects arrived in a bad state at the lab due to a storm. The help offered by the experimenter provoked

is also often used loosely in the case of failures to include a control variable that should have been controlled for, in the sense that its influence on the output is mixed with that of the input variable and therefore contaminates estimates of the relationship between the input and output (Pearl and Mackenzie 2018).

> one of the subjects to say aloud 'How nice? I feel in such a cooperative mood.' This remark was met from the other subjects by laughter and further comments of the same nature (Nikiforakis 2008 p.99).

It would be unthinkable to see this type of problem and explanations in a 'hard' science paper, but they are part of the day-to-day reality for experimental social scientists. One can redesign and redo the experiment as many times as resources allow, but there just will not be enough resources to eliminate every human idiosyncrasy that poses a potential threat to controlled comparison. It would, therefore, seem signally important for understanding the present limitations of the ways in which the experimental method has been implemented in the social and behavioral sciences to gain greater purchase on the range and types of confounds and auxiliary assumptions. To do so we have adopted what might very loosely be called an ethnomethodological approach to the methodological literature, treating it as a resource for identifying what experimenters need to do and to know – even if tacitly – to make their experiments work 'scientifically'.

Methods and methodologies

In this way our study treats disciplinary and interdisciplinary methodological discussions in/across economics and psychology as data. We chose this approach instead of relying on observational data or interviews because we believe these methodological discussions contain insights that cannot be revealed by other methods.

As the next few sections will show, the practices we identify from the methodology literature play enabling roles and make experimentation with human subjects possible. These practices are the substrate on which the experimental method is developed to cope with human subjects, and just like the substrate of a building model, they easily escape the inquiry of observers and interviewers. They are under the radar because so many elaborate manipulation techniques and procedures are built on top of them that they seem peripheral, if not all together invisible – even to most practicing experimentalists, who are usually

preoccupied with learning and developing new manipulations. This is not to say that experimentalists care only about getting things to ‘work’; on the contrary, many are deeply concerned about epistemology and how their methods are shaped by conventions and habitual thinking. Researchers of this stripe often voice their angst and aspiration in methodological discussions, which can be an alternative source for STS researchers seeking insight to the way experimental social scientists work. Looking from the angle of practice, such texts are naturally generated data when communities of practitioners discuss what objectivity means in experimental work and what researchers have to know in order to design and successfully conduct social science experiments.

Taking this approach comes with the challenge of sifting through the broad and diverse interdisciplinary methodological discussions, which can range from statistical issues to philosophical inquiries about the link between theory and practice. Although trained in quantitative social sciences and having done work on behavioral experiments, the first author was driven to explore the methodological and epistemological discussions on social sciences because of her discomfort with the myriad assumptions she found herself having to make during the research process. The data used for this study come mainly from her collection of methodological and epistemological research on how to conduct social science experiments.

In other words, what we rely on here is not a sample designed from the very outset with this particular study in mind; instead, it is a collection accumulated by a practitioner in search of solutions to what she considered to be relevant issues in the epistemology and practice of a series of social science experiments. It is also important to note here, that the first author’s experimental work was mainly in the intersection of economics and social psychology (specifically public goods and rule-breaking), and this limits the coverage of more distant sub-areas such as macro-economics, developmental psychology, and cognitive psychology, just to name a few. Granted, this data collection method does not guarantee

representativeness, much less comprehensiveness, but we believe approaching our research question through the methodological discussions on experimental social sciences can provide new and noteworthy insights into the connections between subject matter, practice, and research findings.

A note on competing terms and terminologies: Neoclassical economics, Behavioral economics, and Conventional Psychology

Just as the term 'natural sciences' includes many sub-fields that are uneven in terms of their mathematical sharpness and predictive precision, experimental social sciences are also made up of heterogenous but often overlapping sub-fields. These blurry edges across sub-fields are to be expected, but they can be a source of confusion partially because the naming of disciplines is often historically contingent. To forestall confusion, this Section provides a brief explanation of three broad categories of experimental social sciences covered in this analysis.

For economists, the term 'economics' usually refers to neoclassical economics. Neoclassical economics assumes that humans are rational-maximizers and it used to solely rely on observational data for empirical testing until Vernon Smith and colleagues introduced the experimental method to the field in the 1960s (Smith 2003). Smith eventually won the 2002 Nobel Memorial Prize in Economic Sciences, and notably he shared the Prize with Danial Kahneman, a psychologist whose work on uncertainty and heuristics with the late Avos Tversky is considered the modern foundation of behavioral economics. Although with the word 'economics' in the label, Kahneman and Tversky's work, and that of new generations of behavioral economists, have their roots in psychology and they share with conventional psychology the focus on context (at least as 'context' defined by their authors). Unfortunately, this history is not well-known outside of these circles, and insiders have a

tendency to keep referring to neoclassical economics as economics and behavioral economics as psychology.

Of course, the umbrella term ‘psychology’ or ‘conventional psychology’ are inevitably impressionistic, too. Experimental psychology is also made up of many heterogeneous and overlapping sub-fields, such as social psychology, cognitive psychology, and neuro-psychology, just to name a few. For our purpose, it suffices to pay attention to the difference between the long-established sub-fields (that often look at human perception and decisions in social contexts) and the younger sub-fields of cognitive and neuro psychology (that tend to study mental phenomena as processes isolated in the human brain). The fact that behavioral economics and conventional psychology share a relatively similar view of human nature translates to greater similarity in methodological practice. In the following discussion, we therefore adopt the conventional division with modified wording: we roughly divide the literature into neoclassical economics and psychology (which includes behavioral economics).

Given the myriad of factors that can influence human behavior, experimental social scientists usually implement controlled comparisons on two fronts. First, both neoclassical economists and psychologists theoretically abstract away factors considered non-essential while at the same time using experimental manipulations to accentuate the model’s key elements. Second, experimenters suppress all other factors (regarded as noise or interference) with several strategies that in general limit permissible subject actions and choices. The next two Sections explore each of these two approaches in succession.

Accentuating Experimental Manipulation: Theoretical Abstraction

Neoclassical economists and psychologists share interests in domains such as group behavior and decision making; they also both rely on theoretical abstraction to amplify their target phenomena against other background factors for experimental manipulation. However, their

approaches to theoretical abstraction differ markedly: neoclassical economists seek to abstract contexts away, whereas psychologists tend to be interested in the way context affects human perception and behavior (Hogarth 2001; Ariely and Norton 2007; Zwick and Budescu 1999; Huettel and Lockhead 2001).

Sub-areas in neoclassical economics are unified under the theoretical assumption that human beings are rational maximizers, which makes it the essential auxiliary assumption in experimental neoclassical economics. Seeing humans as rational maximizers, neoclassical economists believe that people's real preference can best be inferred from their decisions where monetary or material payoff is at stake – putting your money where your mouth is. Based on this premise, if a theoretical model can be translated into a set of decisions and every decision translated into explicit cost and benefit terms, then the results of economics experiments should mirror the performance of the theory in the real world (Smith 1976, 1982). This is why neoclassical economists in general believe that rewards for participants should be pegged to subjects' 'performance' in the lab.

Through recruiting university students with ads highlighting monetary rewards and exacting instructions of payoff calculations in the lab (Böhme 2016; Hertwig and Ortmann 2001), the principle of rational maximization steers subjects' attention every step of the way, abstracting away other human factors such as contextual cues and values. In translating everything into cost/benefit items, neoclassical economists go to great lengths to ensure that participants understand the incentive structure and how it affects their monetary payoff. The instructions are usually written in a highly structured format with numerical examples showing how payoffs are calculated, and the instructions are typically followed by test questions and practice rounds. The test questions and practice rounds give the researchers a chance to identify individuals who are not playing 'rationally' so they can be 'taught' to play the 'right' way (Böhme 2016; Muniesa and Callon 2007). The result is a performance

choreographed according to the researchers' theoretical model (Böhme 2016; Asdal and Cointe 2022).

The rational maximization assumption can also explain other features of neoclassical economics experiments. For example, on the debate over the cost and benefit of deception in experiments, neoclassical economists often argue against deception on the grounds that it will lead subjects to second-guess the true purpose of the experiments, which may undermine the instructions and hinder rational calculation (Bonetti 1998; Hersch 2015). Another example is the temporal structure of neoclassical economics experiments, which may be set up as single- or multiple-round experiments where subjects play just once or play the same game repeatedly. One-shot games offer no learning opportunities and have the simplest information structure. By contrast, when an experimental subject is able to play the same game repeatedly she can try different strategies and factor in future interactions, both of which add information and therefore affect the calculation of payoffs. Neoclassical economists are interested in repeated games mainly because many economic theories assume equilibrium, which requires that people have enough experience and information of the interactions to make rational calculations and optimal decisions (Hertwig and Ortmann 2001; Hyndman et al. 2012).

In contrast to neoclassical economists, many psychologists[4] tend to be interested in the myriad ways in which people's perceptions and behavior can be influenced by (often subtle) contextual factors (Ariely and Norton 2007; Zwick and Budescu 1999). Since context is a catch-all word for things that cannot be comprehensively listed, a universal theory that explains how context affects human perception and behavior does not exist. Probing the

[4] Here the contrast is drawn against cognitive psychology, which seeks to understand the hidden cognitive processes that are assumed to be independent of context.

effects of context on perception and behavior, psychologists therefore tend to model contextual factors from various angles. From this perspective, the difference between neoclassical economics and psychology experiments lies mainly in the degree of theoretical abstraction and what they choose to accentuate through experimental manipulations. Psychologists assume that contextual factors are subject to individual interpretation and cannot be translated into unambiguous cost and benefit items (Ariely and Norton 2007; Zwick and Budescu 1999; Huettel and Lockhead 2001). Under this premise, they build contextual factors directly into their theoretical models and experimental design, using cover stories, planted situational cues, deception, and confederates to simulate real-world situations. Given this focus on context, cover stories, deception and confederates are sometimes considered necessary for studying psychological phenomena that would otherwise be unobservable (Hilton 2001).

For example, a social psychologist studying helping behavior under time pressure has to stage a situation where the subjects encounter a confederate in need and have to decide between helping and being late for their appointment or class. Or, a researcher studying the effect of signs of disorder in the physical environment on whether people litter has to manufacture signs of disorder and plant something that the subjects would want to quickly get rid off. In such situations, the consequences of their action are left for the subjects to define and evaluate, and individuals often perceive the situation differently and therefore take different actions (Henrich, 2001; Henrich, Heine, and Norenzayan 2010).[5] Partly because there is no correct choice in such situations and therefore no 'performance' to speak of in the sense of rational calculation, monetary incentives have been uncommon in psychological

[5] The results are understandably more variable as individual factors play a bigger role in the evaluative process, even when some personal characteristics are statistically controlled for.

experiments (Baron 2001). Instead, psychologists tend to follow a tradition of using course credits to recruit subjects from the population of psychology students. Similarly, repeated games are rarely used in psychology experiments (other than in public-good and learning experiments where repetition may resemble real-world situations) because people's perception of real-world contexts cannot be 'improved' with practice and repetition (Gil-White 2001; Gillies and Rigdon 2001).[6]

Techniques for Suppressing Noise: the role of 'Procedural Duct-tape'

In contrast to their differences in theoretical abstraction (and what they choose to accentuate through manipulation), neoclassical economists and psychologists rely on some common techniques that reduce variation and unexpected or discrepant behavior ('noise'). These techniques - unlike the main design of the experiment routinely set out in the Method section of a research paper - are often glossed-over in publications because they are not purpose-built for individual experiments and are treated as peripheral and non-essential measures for procedural reinforcement. In a sense, they are like duct tape – they are widely used but rarely noticed. We argue that these often-overlooked techniques are important because they perform crucial work that aims to suppress inherent uncertainties and inconsistencies in human behavior, and in doing so they provide the substrate on which the experimental method attempts to cope with the vagaries of malleable and unpredictable human subjects. In what follows we discuss three common noise-suppressing strategies: purifying lab subjects, restricting response space, and excluding contaminated responses.

Purifying Lab Subjects

[6] Cognitive psychology and neuro-psychology do use practice sessions and rehearsals extensively as their focus tends to be (ostensibly) mechanical processes isolated in human brains (Cohn 2008; Martin 2022).

This is a common strategy aiming at homogenizing subject behavior by front-staging lab rules and instructions. As already mentioned in the previous Section, those who participate in economic experiments are usually motivated by monetary rewards and participants of psychology experiments usually take part to fulfill credit requirements or for extra credits (Hertwig and Ortmann 2001). It is fair to say that, for most lab subjects, participation is a form of exchange – their time and cooperation for money or course credits. This exchange is realized through a form of social contract between the researcher and the participants, which requires participants to conform to their roles as subjects and perform as stipulated by the researcher's rules (Böhme 2016; Gozli 2017, 2019).

The social contract starts with the recruitment process. Nowadays recruitment of experiments is usually implemented through centralized recruitment platforms such as ORSEE[7] or the Sona Systems.[8] Upon signing up to these systems, people have to agree to general terms and conditions such as providing a valid mobile phone number and avoiding unexcused absences (those who have accumulated three unexcused absences will have their accounts deactivated). If they agree to these terms and conditions, those signing up enter their contact information into the database of the recruitment system, which then notifies them when new recruitment ads become available. Böhme's study on economic experiments (2016) illustrates the power of the social contract – even before the formal experiment instructions begin, participants already act in a way showing a clear understanding of what to expect if they fail to hold up their end of the bargain. This is also typical of the subjects in psychology experiments as they behave similarly, i.e., willingly conforming to the terms and conditions set up by the recruitment process (Gozli 2017).

[7] This is short for Online Recruitment System for Economic Experiments (http://www.orsee.org/web/).

[8] See https://www.sona-systems.com.

Once inside the lab, the same instructions are read to participants, and a computer interface is used whenever possible so that participants in the same experiment condition receive exactly the same instructions and treatments (Böhme 2016; Guala 2005). The instructions typically make explicit requirements that, in order to receive their payments or course credits, participants must follow rules stipulated by the experimenter throughout the experiment session. In particular, participants usually are banned from using cellphones or talking to each other during the experiment. Also, in experiments with multiple participants working simultaneously, participants are usually seated in cubicles that block their sight of others. These rules and seating arrangements shield the participants from the distraction of cellphones and others. Together with the recruitment process, these procedures and rules gradually and imperceptibly reduce and transform participants to experimental subjects prepared for manipulation and production of controlled information (Muniesa and Callon 2007).

Restricting Response Space

Whereas experimenters can unambiguously channel stimuli (e.g., high- vs. low-stress tasks) to participants, participants can interpret and react to the stimuli in various ways and their responses to open-ended questions can be all over the place. For example, the image of a flying bird may evoke in some people a sense of nature and freedom, in others a sense of lightness, and nothing at all in yet others. Potential responses are many, but standard statistical hypothesis testing requires clearly defined variable attributes and researchers must therefore measure the response variable with closed (rather than open) questions and allow for only a subset of all possible responses (Danziger 1990 pp. 137-138).

Take priming as an example. Priming experiments seek to show that exposure to some concepts automatically and unconsciously biases our decisions/perceptions/feelings in certain directions (Bargh and Melnikoff 2019). In the case that a researcher wants to test if the image

of a flying bird biases people's perception in an upward direction, she can ask subjects to make a sequence of up/down choices, during which an image of a flying bird may be flashed. A positive correlation between seeing the flying bird and choosing 'up' is then taken to indicate that the image of the flying bird primes people in an upward direction. While the image of a flying bird might alternatively prime people to move in a gliding fashion or act in flocks instead of clicking the 'up' arrow key, those options are unavailable since they are beyond the scope of the investigation and are deemed external to the experimental model. By restricting potential responses, this strategy keeps out factors not included in the research model and produces data that look 'clean'.

The widely used Likert scale (those familiar multiple choice questions with options ranging from 'strongly disagree' to 'strongly agree') is another good example. This can be clearly seen in textbook discussions on neutrality in questionnaire design:

> Survey researchers generally agree that more nuanced categories more precisely capture respondents' feelings and that data analysts could combine categories... The 'neither agree nor disagree' option is controversial. Some researchers believe that survey questions should not offer this option; instead, respondents should make a forced choice to indicate their general leanings toward either agreement or disagreement. By contrast, other researchers believe that 'neither agree nor disagree' is a legitimate response for people who are either uninformed about the topic or who genuinely hold ambivalent views (Carr et al. 2018 p. 216).

By forcing subjects to choose from a predetermined set of options, limiting response space plays an important role in extracting (seeming) clarity out of fuzziness and achieving the appearance of objectivity and standardization (more on this in the next Section).

Excluding Contaminated Responses

The third noise-suppressing strategy works by excluding observations whose responses are considered 'contaminated'. The logic of the exclusion decision is based on backward induction: if individual participants respond in unusual and 'non-sensical' way, their performance must have been marred by misunderstanding of the tasks, error in execution, or

something extraneous. Put differently, 'there must be a basic level of competence required of subjects' (Martin 2022 pp. 180-181), but that level is often determined *post hoc*.

For instance, in a study on moral licensing/cleansing[9], Ho and colleagues (2016) exclude several observations 'because at least one component of their carbon footprint was much greater than the rest of the sample, often an order of magnitude more. These observations were unrealistically high values, appearing to be incorrectly entered responses…' Similarly, in testing how worry affects people's willingness-to-pay for insurance under uncertainty, Schade, Kunreuther, and Koellinger (2012) exclude some observations 'because participants bid more for insurance than 10 000 times the expected loss', which is the inverse of the odds at which they may lose their endowed valuables according to the experiment instructions. The taken-for-granted assumption here is that no rational people would pay more than the present value of their endowed valuables to insure its future. In other words, such responses are taken to imply irrationality, which justifies the exclusion of the participants' data.

In psychology, exclusion is often used alongside manipulation checks, which are most valuable when participant awareness of the stimulus (the input variable) is required but in doubt. Manipulation checks provide measures of participants' perception of the stimulus and they are used to assess whether the stimulus was received by participants as designed (Sigall and Mills 1998). Manipulation checks usually take the form of verbal questions and are conducted right after the administration of the stimulus and/or before the measurement of the output variable. They can be used in the design stage to help researchers assess whether the

[9] This is the idea that those who are told that they do better/worse than the average person in one domain, such as carbon footprint, will consequently behave in the opposite way in another domain such as buying green energy.

stimulus needs to be enhanced. Alternatively, manipulation checks are often used in the data analysis stage to exclude the data of participants who fail to notice the stimulus (e.g., Bahns and Crandall 2013).

When manipulation checks are used to make design decisions, it is a form of accentuating the stimulus, but when manipulation checks are used in the data analysis stage to exclude observations from analysis, the logic is similar to the notion of contaminated input as described above: although misunderstanding or execution errors may not be the reasons for participants' failure to notice the manipulation, there must be something that prevented them from taking notice of the stimulus, and that 'something' forms the grounds for excluding these participants' responses because it is an unaccounted-for, extraneous factor. In this sense, exclusion of those participants who fail manipulation checks works as a protection against potential confounds that the conceptual model does not anticipate.

Human Complexity, Experimenters' Regress, and the Reproducibility Crisis

The two preceding Sections have described the two-pronged approach – theoretical abstraction and suppressing noise – that experimenters follow in order to extract their target phenomena from the unpredictable situations in the social science lab. At the conceptual level, theoretical abstraction guides experimental design, which foregrounds the key features of the theoretical model and directs subjects' attention to them. Meanwhile, seemingly mundane and inconspicuous administrative procedures and implementation techniques actually do important work in homogenizing subject responses and reducing variation in observed outcomes. Taken together, these practices transform lively participants into cooperative workers who produce information conforming to the research community's standards of objectivity. In addition to these general patterns, methodological discussions on social science experiments also reveal important insights on the limits of these practices, and they are most visible when established results are challenged and fire is exchanged between

opposing camps. This Section illustrates these limits with high profile debates and controversies.

Being the foundation of neoclassical economics, the rational maximization assumption nevertheless has long faced criticisms from other disciplines, and it has been a topic of ongoing debate among neoclassical economists and psychologists (Berg and Gigerenzer 2007; Hogdson 2001). Challenges often come in the form of empirical evidence showing the prevalence of 'irrational' decision-making. For example, a large body of behavioral economics literature on the 'endowment effect', where people place a higher value on an item that belongs to them than on an otherwise identical item (Ericson and Fuster 2014; Knetsch 1989), directly questions the notion of 'rational' agents who make decisions based on market value.[10] This pattern is usually demonstrated with one of two experimental paradigms: the exchange paradigm and the valuation paradigm. The exchange paradigm randomly gives subjects one of two items of the same market value (e.g., a mug vs. a pen) to begin with. A few minutes later subjects are given the chance to trade their given item with the experimenter for the other item. In the valuation paradigm, half of the subjects are randomly assigned as sellers and endowed with an item (such as a mug) and the other half assigned as buyers. The experimenter then elicits buyers' willingness-to-pay and sellers' willingness-to-accept for the item for comparison. The patterns observed in the exchange paradigm show that most subjects decide to keep their initial endowment, and the valuation paradigm tends to find that sellers see greater value in the item than buyers do. Both suggest that ownership affects perception of value.

Unsurprisingly, neoclassical economists counter the challenge on methodological grounds. Plott and Zeiler (2005) show that, under the valuation paradigm, the valuation gap

[10] This undermines presumed efficiency of trade and has important economic and legal implications.

between owners and buyers disappears when a few elements are added: elaborate explanation on the value elicitation mechanism, illustrations of the optimal strategy, and practice rounds. They therefore argue that the original results were artefacts arising from experimental subjects' misconception of the task (e.g., confusion with the common 'buy low, sell high' heuristics). In another study targeting the exchange paradigm, Plott and Zeiler (2007) eliminate the pattern of the endowment effect by changing the wording and emphasizing to subjects that the items are randomly distributed. This lends weight to their argument that the subjects in the earlier exchange studies may have mistakenly inferred that the 'gift' (used in the original instructions) was chosen for them and therefore might be 'better' or should not be traded in. This pair of studies triggered another round of search for 'procedural artefacts' (for a review see Ericson and Fuster 2014) and have been on the receiving end of similar criticism as well. Despite the manipulation-accentuating and subject-purifying strategies at their disposal, decades later experimentalists still dispute about the 'endowment effect', producing a dizzying array of additional parameters that may explain its presence/absence.

Other noise-suppressing strategies have their own limits, too. Although restricting subjects' response space and/or excluding contaminated observations are powerful ways to reduce noise, this is often achieved through restricting responses to a very narrow set and defining most factors as external (Danziger 1990; Gozli 2017, 2019). These restrictions qualify the findings and often severely limit their generalizability. The Nobel Prize-winning experimental work on prospect theory (Kahneman and Tversky 1979; Tversky and Kahneman 1992) serves as an example. Prospect theory suggests that people are risk-averse when facing gains but risk-seeking when facing losses. Tversky and Kahneman demonstrated this pattern with a series of experiments where subjects are required to choose between two bets with the same expected values but with different risk profiles. For example, in one bet the subjects have an 90% chance of winning (or losing) $20 and a 10% chance of winning (or

losing) $0. And in another best the subject may win (or lose) $90 with a 20% chance and $0 with an 80% chance. The expected gain (or loss) is $18 in both bets. Given the same expected gains, subjects tend to choose the bet with a greater possibility of earning a moderate amount (90% chance of winning $20) over the one with a low possibility of making a larger amount of money (20% chance of winning $90). Conversely, given the same expected losses, subjects are more likely to choose the bet with a low possibility of losing a larger amount of money (20% chance of losing $90) over the bet with a greater possibility of losing a moderate amount (90% chance of losing $20). These are fairly robust findings; however, they do not account for the nature of the experimental tasks. As the literature on preference reversal (Slovic 1995; Tversky, Slovic, and Kahneman 1990) shows, if the task asks subjects to price those two bets (instead of choosing), then subjects behave differently and are willing to pay more for the bet that offers a higher payoff at a lower possibility. That is, the outcome is contingent on the response format.

In a revealing comment on the growing list of parameters generated by the debate over monetary incentive on experimental results, Nobel Laureate Vernon L. Smith graciously acknowledges:

> The theory forever lags behind the empirical results, yielding what Lakatos calls 'miserable degenerating research programmes'… This undesirable state is a consequence of the rhetorical commitment to falsificationist/predictive criteria. Why should we believe that we can construct falsifiable, predictive models by abstract thought alone? If we have learned anything in 50 years of experimental economics it is that real people do not solve strategic decision problems by thinking about them the way we do. In fact, we do not solve our own decision problems this way, except in our publications. There isn't time, it's cognitively too costly; and if the problem has really high stakes (the vast majority of daily decisions have low stakes), we hire professionals… Our task should be to modify theory in the light of evidence, and aspire to encompass suspected auxiliary hypotheses (stakes, subject sophistication) explicitly into the theory to motivate new tests. (Smith 2001 p. 428)

Smith then changes track and ends his comment with the potential of new technologies such as brain imaging. Although seemingly sudden, this turn makes a lot of sense if seen as a reflection of the aspirational nature of 'modify[ing] theory in the light of evidence and aspir[ing] to encompass suspected auxiliary hypotheses … explicitly into the theory…'

Similar problems plague psychology as well. Derksen's detailed account of the controversy over priming (2017) illustrates how various resources are mobilized by those defending and challenging the earlier findings. When pressed to put forward auxiliary assumptions for replication studies, experts eventually turn to 'insufficient theoretical understanding' to explain replication failures. Similar to the example of endowment effect above, efforts to improve priming theory result in an ever-expanding list of variables, leaving the once powerful concept much diminished.

Reviewing 13 long-running controversies in psychology, Greenwald (2012) concludes that 'publications that were treated by one side as crucial opposition-falsifying findings were generally greeted by the opposed side as conceptually or empirically flawed efforts.' Essentially, this points to the difficulty of meeting auxiliary assumptions and the interpretive aspect of replication (Freese and Peterson 2017). As psychologists are at the forefront of the recent replication debate, a few controversies in psychology became polemical and the highly charged exchanges put the impossibility of bullet-proofing experimental results on display. Vehement disagreements on what counts as successful replications have been a feature in another two recent controversies, one over ego depletion (Baumeister 2019; Baumeister and Vohs 2016; Hagger and Chatzisarantis 2016), and the other over power posing (Credé 2019; Cuddy, Schultz, and Fosse 2018). In both cases, high profile findings came under scrutiny and failed to replicate, and the authors of the original studies defended their findings by arguing either that the replication studies fail to preserve some key features of the original tasks/procedures, or that the replication studies included in meta-analysis do not match their original conceptual framework. In these on-going debates, auxiliary assumptions become the focal point, and the debates provide vivid illustrations of the infinite experimenters' regress. As long as researchers are unable to surgically isolate any disputed factors, the door is always

open to arguing on the grounds of heretofore unarticulated auxiliary assumptions, making any final consensus hard to reach.

Conclusion

Physics and younger sciences (particularly the social sciences) have long existed in a kind of epistemological tension. The extraordinary breakthroughs in physics have for some time set a benchmark for the advancement of knowledge as based on a process that is cumulative, incremental, technically applicable, replicable, and so on. It is understandable, therefore, that the dominant method of physics should be adopted by younger sciences, partly in the hope of matching the successes of physics and partly in the hope of matching its prestige.

This indiscriminate transfer of method becomes problematic when it obscures inherent mismatches between phenomena and research methods. Physics is concerned pre-eminently with objects or phenomena that are broadly insensate and therefore more amenable to experimental manipulations, observation and calculation. As Nelson points out,

> given the questions physicists ask about them, by and large all electrons have the same basic characteristics, and parametric specification of a few variables suffices to identify their variation that is relevant to physicists. While there may be a stochastic element, electrons behave similarly under similar conditions, at least regarding the behaviors and conditions that physicists have chosen to study. And most of the conditions and forces affecting the state and behavior of electrons that physicists study, for example the strength of a magnetic field, can be assessed and described in a general way, and their effects explored under controlled conditions that shield the context from variation in other factors (Nelson 2016, p. 1697).

In contrast to inanimate matters, living systems (as the subject matter of biology and human sciences) feature greater variation and heterogeneity, complex intra-system hierarchies, inseparability (of effects), and emerging properties (Nelson 2016; Mayr 1985). Even within the same species, living organisms come in all shapes and sizes; and together with cross-breeding such heterogeneity inevitably results in fuzzy categorization and measurement. Moreover,

> [c]omplexity in living systems exists at every level from the nucleus (with its DNA program), to the cell, to any organ system (like kidney, liver, or brain), to the individual, the ecosystem, or the society. Living systems are invariably characterized by elaborate feedback mechanisms, unknown in their precision and complexity in any inanimate system. They have the capacity to respond to external

> stimuli, the capacity for metabolism (binding or release of energy), and the capacity to grow and to differentiate (Mayr 1982 p.53).

Studying humans poses even greater challenges because human subjects also have consciousness and social agency, and as Danziger reflected in *Constructing the Subject* 'the outcome of the investigation was the product of a social interaction in a role system whose structure was intimately connected with the way in which the object of investigation had been defined' (Danziger 1990 p. 31). In other words, the fact that most experiments involve 'scripts' designed on the one hand by the experimenter but dependent on the other hand on experimental subjects who must respond to their manipulation in the experimental context makes the outcome of the experiment – its results – a co-construction flowing from an asymmetrical relationship. At the same time, however, although we have described the experimental situation in the course of this article in ways that show the particular axes of asymmetry, as if power lies with the experimenter and not with the experimental subject, the experimentee, there is an important sense in which both are in the last analysis 'subjects'. For the force field of power and control does not flow simply on a one-dimensional axis from the experimenter to the experimentee. The experimenter is in their own turn, and in an important sense, a subject, subjected to the protocols and manoeuvres of the subfield that constitutes his/her community of practice, required to make theoretical abstractions, required to purify his/her experimental subject and required to restrict their response space in order to secure experimental results. Ultimately, the kinds of questions that can be studied about human behavior are constrained by the social context of the data collection process, be this experiments, surveys, or observational studies.

Since the nature of phenomena determines the kind of questions that can be asked and the methods that can be fruitfully employed to study them, we support Longino's call for a

pluralistic approach (2013)[11] that recognizes and embraces the complexity and fluidity of human behavior. In addition to the experimental method, there are many choices (including some associated with other disciplines only for historical reasons). For example, whereas some linguists focus on interactions and tap into naturally occurring data from daily contexts, others draw on introspection and intuition (as in generative linguistics). Many important contributions are also made by sociologists and anthropologists working with ethnographic data. In principle, we should encourage scholarship from different angles and different methodological traditions, as long as the match between the research question and the chosen method is a good one. This is because, in the bigger picture, the real problem underlying the methodological difficulties of the social sciences might not be methods so much as institutional pressures on applied researchers and the attendant lack of incentives for mindful research and evaluation. This requires more education on the philosophical side of science, as well as a philosophy of science that does not privilege studies of natural sciences.

With inadequate understanding of the way science works, our society is enamoured by scientificity,[12] to the point that even researchers often forget about Statistician George Box's famous quote 'Essentially, all models are wrong, but some are useful' (Box 1987 p.424),[13] which naturally extends to theories and empirical studies that seek to test theories. Researchers often blindly chase realism by continuously adding variables, all the while

[11] Longino (2013) surveys research approaches to studying human behavior and demonstrates ways in which inseparability of effects and incommensurable approaches threaten research validity and makes policy research problematic.

[12] Though note the hostile reaction of some populist politicians to the role of 'experts' in the determination of policy.

[13] Compare the sociolinguist Labov: "Formalisation is a useful procedure even when it is wrong: it sharpens our questions and promotes the search for answers" (Labov 1972 p. 61).

forgetting that the best we can hope for is models and methods that work – for our specific purposes. Such mindless practices are further exacerbated by widespread misconceptions and abuse of statistics among applied researchers, which create a large body of work that cannot withstand scrutiny (Gelman 2022; Gelman and Loken 2014). Given the scope of the problem, there may be no easy solutions to our methodological difficulties – after all, all models/theories are subject to revision. However, maybe we can ask what makes some empirical studies more useful than others. We believe that one condition for useful research is thoughtful matching between the specific theoretical questions and the empirical approaches taken. This 'match' underpins research quality and is likely to have a substantial impact on replicability. When a method is forced on a problem that is not defined in a compatible way, validity suffers and there is little hope of achieving the level of mathematical sharpness and predictive precision of physics.

Asking applied researchers to come out of their comfort zone and embrace methodological plurality and uncertainty is no doubt difficult. In the modern day, societies are constantly struggling with a scalability issue where the sorting mechanisms are always sub-optimal, and most people are just getting by with out-dated/partial information and do not have time to think more critically. For promoting mindful research, a literature that might be relevant here is work on the 'prediction markets', which has been used to test the replicability of research findings with some interesting initial findings (Camerer et al. 2018; Dreber et al. 2015). The concept of prediction markets, or more broadly 'idea futures', is the brainchild of Robin Hanson (1995). Seeking a mechanism to incentivise better evaluation of research, Hanson proposes to let people bet on research outputs so academic insiders have to 'put their money where their mouth is.' (Interestingly, this idea was first published in this very Journal in 1995.) If community insiders can do reasonably well predicting replicability, then there is probably no need to write off experimental method (or any other method for that matter)

wholesale for a discipline, which is typically diverse and heterogeneous to begin with. Prediction markets may provide a more robust mechanism that can withstand the deluge of increased number of publications – provided that we are cautious and not let the idea be used mindlessly, as that is what humans do when we are pressed for time. In the end, the illusion of research productivity and efficiency may prove to be more detrimental to social sciences than imperfect methods.

References

Ariely, Dan, and Michael I Norton. 2007. "Psychology and Experimental Economics: A Gap in Abstraction." *Current Directions in Psychological Science* 16 (6): 336–339.

Asdal, Kristin and Béatrice Cointe. 2022. "Writing Good Economics: How Texts 'On the Move' Perform the Lab and Discipline of Experimental Economics." *Social Studies of Science*. Advance online publication. doi: 030631272210796.

Bahns, Angela, and Christian Crandall. 2013. "The Opposite of Backlash: High-SDO People Show Enhanced Tolerance When Gay People Pose Little Threat." *European Journal of Social Psychology* 43(4): 286–291.

Bargh, John, and David Melnikoff. 2019. "Does Physical Warmth Prime Social Warmth?" *Social Psychology* 50(3): 207–210

Baron, Jonathan. 2001. "Purposes and Methods." *Behavioral and Brain Sciences* 24 (3): 403.

Baumeister, Roy. 2019. "Self-control, Ego Depletion, and Social Psychology's Replication Crisis." https://psyarxiv.com/uf3cn/download?format=pdf.

Baumeister, Roy, and Kathleen Vohs. 2016. "Misguided Effort with Elusive Implications." *Perspectives on Psychological Science* 11 (4): 574–575.

Berg, Nathan, and Gerd Gigerenzer. 2007. "Psychology Implies Paternalism? Bounded Rationality May Reduce the Rationale to Regulate Risk-taking." *Social Choice and Welfare* 28 (2): 337-359.

Birenbaum, Menucha, and Kikumi K. Tatsuoka. 1987. "Open-Ended Versus Multiple-Choice Response Formats—It Does Make a Difference for Diagnostic Purposes." *Applied Psychological Measurement* 11 (4): 385–395.

Box, George, and Norman Draper. 1987. *Empirical Model-Building and Response Surfaces*. New York, NY: Wiley.

Böhme, Juliane. 2016. "'Doing' Laboratory Experiments: An Ethnomethodological Study of the Performative Practice in Behavioral Economic Research." In *Enacting Dismal Science*, edited by Ivan Boldyrev and Ekaterina Svetlova, 87–108. New York: Palgrave Macmillan.

Bonetti, Shane. 1998. "Experimental Economics and Deception." *Journal of Economic Psychology* 19 (3): 377–395.

Camerer, Colin. F., Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler et al. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015." *Nature Human Behaviour* 2: 637–644.

Carr, Deborah, Elizabeth Heger Boyle, Benjamin Cornwell, Shelley Correll, Robert Crosnoe, Jeremy Freese, and Mary Waters. 2018. *The Art and Science of Social Research*. New York, NY: WW Norton & Company.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Oxford, UK: Oxford University Press.

Cohn, Simon. 2008. "Making Objective Facts from Intimate Relations: The Case of Neuroscience and its Entanglements with Volunteers." *History of the Human Sciences* 21 (4): 86-103.

Collins, Harry. 1985. *Changing Order: Replication and Induction in Scientific Practice.* London, UK: Sage Publications.

Créde, Marcus. 2019. "A Negative Effect of a Contractive Pose is Not Evidence for the Positive Effect of an Expansive Pose: Comment on Cuddy, Schultz, and Fosse (2018)." *Meta-Psychology* 3. https://doi.org/10.15626/MP.2019.1723.

Cuddy, Amy, Jack Schultz, and Nathan Fosse. 2018. "P-Curving a More Comprehensive Body of Research on Postural Feedback Reveals Clear Evidential Value for Power-

Posing Effects: Reply to Simmons and Simonsohn (2017)." *Psychological Science* 29 (4): 656-666.

Danziger, Kurt. 1990. *Constructing the Subject: Historical Origins of Psychological Research*. Cambridge, UK: Cambridge University Press.

Derksen, Maarten. 2017. *Histories of Human Engineering: Tact and Technology*. Cambridge, UK: Cambridge University Press.

Dreber, Anna, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, et al. 2015. "Using Prediction Markets to Estimate the Reproducibility of Scientific Research." *Proceedings of the National Academy of Sciences* 112 (50): 15343–47. doi:10.1073/pnas.1516179112.

Ericson, Keith M. Marzilli, and Andreas Fuster. 2014. "The Endowment Effect." *Annual Review of Economics* 6 (1): 1–25. doi:10.1146/annurev-economics-080213-041320.

Freese, Jeremy and David Peterson. 2017. "Replication in Social Science." *Annual Review of Sociology* 43 (1): 1–19. doi:10.1146/annurev-soc-060116-053450.

Gelman, Andrew. 2022. "Criticism as Asynchronous Collaboration: An Example from Social Science Research." *Stat* 11 (1). doi:10.1002/sta4.464.

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102 (6): 460. doi:10.1511/2014.111.460.

Gibson, Stephen. 2019. *Arguing, Obeying and Defying: A Rhetorical Perspective on Stanley Milgram's Obedience Experiments*. Cambridge, UK: Cambridge University Press.

Gil-White, Francisco J. 2001. "A Good Experiment of Choice Behavior is a Good Caricature of a Real Situation." *Behavioral and Brain Sciences* 24 (3): 409-410.

Gillies, Anthony S., and Mary Rigdon. 2001. "Theory-testing Experiments in the Economics Laboratory." *Behavioral and Brain Sciences* 24 (3): 410-411.

Gozli, Davood. 2017. “Behaviour Versus Performance: The Veiled Commitment of Experimental Psychology.” *Theory & Psychology* 27 (6): 741–758.

Gozli, Davood. 2019. *Experimental Psychology and Human Agency*. Cham, Switzerland: Springer.

Greenwald, Anthony G. 2012. There Is Nothing So Theoretical as a Good Method. *Perspectives on Psychological Science* 7 (2): 99–108.

Guala, Francesco. 2005. *The Methodology of Experimental Economics*. Cambridge, UK: Cambridge University Press.

Hacking, Ian. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science.* Cambridge, UK: Cambridge University Press.

Hagger, Martin, and Nikos Chatzisarantis. 2016. “Commentary: Misguided Effort with Elusive Implications, and Sifting Signal from Noise with Replication Science.” *Frontiers in Psychology* 7: 621.

Henrich, Joseph. 2001. “Challenges for Everyone: Real People, Deception, One-Shot Games, Social Learning, and Computers.” *Behavioral and Brain Sciences* 24 (3): 414-415.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. “The Weirdest People in the World?" *Behavioral and Brain Sciences* 33 (2-3): 61-83.

Hanson, Robin. 1995. “Could Gambling Save Science? Encouraging an Honest Consensus.” *Social Epistemology* 9 (1): 3–33. doi:10.1080/02691729508578768.

Hersch, Gil. 2015. “Experimental Economics’ Inconsistent Ban on Deception. Studies.” *History and Philosophy of Science* 52: 13–19.

Hertwig, Ralph and Andreas Ortmann. 2001. “Experimental Practices in Economics: A Methodological Challenge for Psychologists?” *Behavioral and Brain Sciences* 24 (3): 383–403.

Hilton, Denis J. “Is the Challenge for Psychologists to Return to Behaviourism?.” *Behavioral and Brain Sciences* 24 (3): 415.

Ho, Benjamin, John Taber, Gregory Poe, and Antonio Bento. 2016. “The Effects of Moral Licensing and Moral Cleansing in Contingent Valuation and Laboratory Experiments on the Demand to Reduce Externalities.” *Environmental and Resource Economics* 64 (2): 317–40.

Hodgson, Geoffrey M. 2001. *How Economics Forgot History: The Problem of Historical Specificity in Social Science*. London, UK: Routledge.

Hogarth, Robin M. 2001. “To What are We Trying to Generalize?.” *Behavioral and Brain Sciences* 24 (3): 416-417.

Huettel, Scott A., and Gregory Lockhead. 2001. “Variability is not Uniformly Bad: The Practices of Psychologists Generate Research Questions.” *Behavioral and Brain Sciences* 24 (3): 418-419.

Hyndman, Kyle, Erkut Y. Ozbay, Andrew Schotter and Wolf Ze’ev Ehrblatt. 2012. “Convergence: An Experimental Study of Teaching and Learning in Repeated Games.” *Journal of the European Economic Association* 10 (3): 573–604.

Kahneman, Daniel, and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision under Risk.” *Econometrica* 47 (2): 263-292.

Knetsch, Jack L. 1989. “The Endowment Effect and Evidence of Nonreversible Indifference Curves.” *American Economic Review* 79:1277–1284

Labov, William. 1972. “The Study of Language in its Social Context.” In *Sociolinguistic Patterns*, edited by William Labov, 183-259. Philadelphia, PA: University of Philadelphia Press.

Longino, Helen E. 2013. *Studying Human Behavior: How Scientists Investigate Aggression and Sexuality*. Chicago, IL: The University of Chicago Press.

Martin, Emily. 2022. *Experiments of the Mind: From the Cognitive Psychology Lab to the World of Facebook and Twitter*. Princeton, NJ: Princeton University Press.

Mayr, Ernst. 1985. "How Biology Differs from the Physical Sciences." In *Evolution at a Crossroads: The New Biology and the New Philosophy of Science*, edited by David Depew and Bruce Weber, 43-63. Cambridge, MA: A Bradford Book.

Mayr, Ernst. 1982. *The growth of biological thought: Diversity, Evolution, and Inheritance.* Cambridge, MA: Belknap Press of Harvard University Press.

Morawski, Jill. 2015. "Epistemological Dizziness in the Psychology Laboratory: Lively Subjects, Anxious Experimenters, and Experimental Relations, 1950–1970." *Isis* 106 (3): 567–597.

Morrison, Hazel, Shannon McBriar, Hilary Powell, Jesse Proudfoot, Steven Stanley, Des Fitzgerald, and Felicity Callard. 2019. "What is a Psychological Task? The Operational Pliability of 'Task' in Psychological Laboratory Experimentation." *Engaging Science, Technology, and Society* 5: 61–85.

Muniesa, Fabian, and Michel Callon. 2007. "Economic Experiments and the Construction of Markets." In *Do Economists Make Markets? On the Performativity of Economics*, edited by Donald Mackenzie, Fabian Muniesa, and Lucia Siu, 163–189. Princeton, NJ: Princeton University Press.

Nelson, Richard R. 2016. "The Sciences Are Different and the Differences Matter." *Research Policy* 45 (9): 1692–1701. doi:10.1016/j.respol.2015.05.014.

Nikiforakis, Nikos. 2008. "Punishment and Counter-Punishment in Public Good Games: Can We Really Govern Ourselves?" *Journal of Public Economics* 92 (1–2): 91–112.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. New York, NY: Basic Books.

Peterson, David. 2015. “All that is Solid Bench-Building at the Frontiers of Two Experimental Sciences.” *American Sociological Review* 80 (6): 1201-1225.

Peterson, David. 2016. “The Baby Factory.” *Socius* 2: doi:10.1177/2378023115625071.

Pickering, Andrew. 1995. *The Mangle of Practice: Time, Agency, and Science*. Chicago, IL: The University of Chicago Press.

Plott, Charles R., and Kathryn Zeiler. 2005. “The Willingness to Pay–Willingness to Accept Gap, the ‘Endowment Effect’, Subject Misconceptions, and Experimental Procedures for Eliciting Valuations.” *American Economic Review* 95:530–545.

Plott, Charles R., and Kathryn Zeiler. 2007. “Exchange Asymmetries Incorrectly Interpreted as Evidence of Endowment Effect Theory and Prospect Theory?” *American Economic Review* 97:1449–1466.

Rosenthal, Robert, and Ralph L. Rosnow. 2009. *Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnow's Classic Books*. Oxford, UK: Oxford University Press.

Schade, Christian, Howard Kunreuther, and Philipp Koellinger. 2012. “Protecting Against Low-Probability Disasters: The Role of Worry.” *Journal of Behavioral Decision Making* 25 (5): 534–43.

Sigall, Harold, and Judson Mills. 1998. “Measures of Independent Variables and Mediators are Useful in Social Psychology Experiments: But are They Necessary?” *Personality and Social Psychology Review* 2 (3): 218–226.

Slovic, Paul. 1995. “The Construction of Preference.” *American Psychologist* 50 (5): 364–371.

Sismondo, Sergio. 2010. *An Introduction to Science and Technology Studies*. West Sussex, UK: Wiley-Backwell.

Smith, Vernon. 1976. “Experimental Economics: Induced Value Theory.” *American Economic Association* 66 (2): 274–279.

Smith, Vernon. 1982. “Microeconomic Systems as an Experimental Science.” *The American Economic Review* 72 (5): 923–955.

Smith, Vernon L. 2001. “From Old Issues to New Directions in Experimental Psychology and Economics.” *Behavioral and Brain Sciences* 24 (3): 428-429.

Smith, Vernon L. 2003. “Constructivist and Ecological Rationality in Economics.” *American Economic Review* 93 (3): 465-508.

Stam, Henderikus J., H. Lorraine Radtke, and Ian Lubek. 2000. Strains in Experimental Social Psychology: A Textual Analysis of the Development of Experimentation in Social Psychology." *Journal of the History of the Behavioral Sciences* 36 (4): 365-382.

Teo, Thomas. 2010. “What is Epistemological Violence in the Empirical Social Sciences?” *Social and Personality Psychology Compass* 4 (5): 295-303.

Ting, Carol, and Richard Fitzgerald. 2020. “The Work to Make an Experiment Work.” *International Journal of Social Research Methodology* 23 (3): 329-345.

Tversky, Amos, and Daniel Kahneman. 1992. “Advances in Prospect Theory: Cumulative Representation of Uncertainty.” *Journal of Risk and Uncertainty* 5 (4): 297–323.

Tversky, Amos, Paul Slovic, and Daniel Kahneman. 1990. “The Causes of Preference Reversal.” *The American Economic Review* 80 (1): 204–217.

Zwick, Rami, Ido Erev, and David Budescu. 1999. “The Psychological and Economical Perspectives on Human Decisions in Social and Interactive Contexts.” In *Games and Human Behavior: Essays in Honor of Amnon Rapoport*, edited by David Budescu, Ido Erev, and Rami Zwick. New York, NY: Erlbaum.