# Evaluating Experimental Designs*

**Erik Snowberg**
University of Utah,
CESifo, and NBER
snowberg@eccles.utah.edu
eriksnowberg.com

**Leeat Yariv**
Princeton University,
CEPR, CESifo, and NBER
lyariv@princeton.edu
lyariv.com

April 11, 2025

**Abstract**

This introductory chapter outlines key criteria for evaluating experimental measures, and connects these criteria to the selection of experimental parameters across various contexts. We aim for this chapter to serve as a valuable framework for assessing the different measures, elicitations, and designs explored throughout the handbook.

*We are indebted to Isabelle Brocas, Colin Camerer, Jonathan Chapman, Dietmar Fehr, Simon Gächter, Andrej Gill, Andreas Grunewald, Holger Herz, Florian Hett, Michael Kosfeld, Marta Kozakiewicz, Erin Krupka, Noemi Peter, David Poensgen, Hannah Schildberg-Hörisch, Zahra Sharafi, Charlie Sprenger, Stefan Trautmann, and Sevgi Yuksel for useful discussions.

# 1 Introduction

This handbook is all about measurement: of beliefs, political tendencies, social interactions, and more. What makes one measure or elicitation better than another? What are criteria for assessing various experimental design choices? In this introductory chapter, we explore several aspects of measurement and elicitation quality. Measurement challenges tend to depend on whether the purpose of the experiment is measuring population attributes or testing theories. This chapter aims to articulate these challenges, and offer a framework and vocabulary for discussing, and, hopefully, addressing them. The criteria we outline do not always lead to the same ranking of measures, prompting trade-offs. Articulating these trade-offs will provide a useful lens through which to evaluate the various measures, elicitations, and designs discussed throughout the handbook.

There is a range of research that falls under the experimental banner. Broadly speaking, this research differs according to whether or not there is experimental variation. Elicitations without experimental variation tend to draw on methods developed in economics laboratories, which is why they have inherited the label of "experimental." These studies are usually interested in measuring the distribution of some behavior or preference either in, or across, populations (see, for example, Snowberg and Yariv, 2021). Work that uses experimental variation, on the other hand, is often interested in either testing a theory, or distinguishing between competing theories.[1]

For studies focused on measuring individual or population-level differences in behaviors, the primary challenge lies in developing accurate measures of those behaviors. It is typically impossible to gauge the accuracy of such measures, even at great cost: how can one determine the *true* risk attitudes, social preferences, or political affinities of an individual? Consequently, the design and justification of measures often, perhaps unknowingly, revolve around various criteria intended to bring them closer to accuracy. The first part of this

[1]We use "theory" broadly. Experimental work is, at times, not directed by an existing model, but rather aims to inform future models by identifying new behavioral features: heuristics and biases (Tversky and Kahneman, 1974), altruism (Güth et al., 1982), and the like. Even those types of experiments tend to compare a clear narrative of behavior to the standard predictions of normative rationality, which we view as an exercise driven by theory.

chapter introduces key concepts that we view as valuable for evaluating economic measures. These concepts are drawn and adapted from classical measurement theory and its extension to the social sciences, particularly psychology. Building on this foundation, we discuss how these principles inform and shape several ongoing debates within the academic community.

For experiments that test theories, the primary challenge lies in choosing optimal parameter values. This is complicated by the fact that, while the theories tested may be well specified, specific parameters may lead to more measurement noise. Thus, the selection of outcome measures and parameters is frequently based on an experimenter's experience and hunches. In the second part of this chapter, we formalize the decision problems experimenters encounter when selecting parameters to achieve various experimental objectives. We hope that by routinely applying the criteria outlined in the first part of the chapter to develop better outcome measures, experimenters can combine those measures with the framework presented in the second part to more easily communicate their design decisions.

The experiments discussed across both parts of this chapter share some common ground. Experimental variation may be employed to assess individual or population-level behavior: for example, experimenters can manipulate the parameters of dictator or trust games (Engel, 2011; Johnson and Mislin, 2011). Additionally, a single set of parameters can sometimes suffice to falsify a theory. For instance, the widespread giving observed in dictator games, regardless of parameter values, has been used to challenge theories that exclude social preferences. Consequently, the concepts explored in both sections may be relevant to all experimentalists, whether their focus is on measuring behavior or testing theoretical predictions.

## 2 Evaluating Measures

This section describes six criteria for evaluating a measure in experimental or behavioral economics: construct validity, responsiveness, predictive validity, cost, reliability, and stability. A fundamental, and perhaps obvious, criterion we do not discuss is *accuracy*: how well

a measure reflects the true underlying value of what it aims to assess. Unfortunately, when it comes to human behavior, perfect measurement is rarely possible, even with significant resources. Consequently, evaluation of measures in terms of accuracy is usually unattainable in the social sciences. The criteria presented below offer various compromises, and choosing among them involves a subjective element. Nonetheless, there is a natural ranking of some of these criteria, which guides the order in which they are presented.

The list of criteria we describe is not exhaustive; we have distilled it from various fields to focus on those we believe are the most pertinent to experimental and behavioral economics. We also adapt terminology to align with what appear to be the priorities of economists. This adaptation is partly justified by the lack of universal definitions for many of the criteria across psychology, statistics, and the natural sciences. In the following section, we use these criteria to reframe several current discussions in the experimental and behavioral literature, highlighting the trade-offs they present.

## 2.1 Construct Validity

*Construct validity* refers to whether a measure theoretically captures what it purports to represent.

Construct validity can be evaluated by ascertaining whether or not the measure respects the comparative statics that the construct it represents should exhibit. For example, for expected utility maximization with utility functions that are universally concave or universally convex over the range of payoffs, certainty equivalents of lotteries should vary with mean preserving spreads of a lottery—decreasing if concave, increasing if convex. Similarly, for social preferences—say, those reflecting a taste for equal outcomes—a split of a fixed budget should respond monotonically to the exchange rates corresponding to each participant. Put another way, data generated by a measure with high construct validity can falsify its theoretical premise.[2]

[2]If data from a measure with high construct validity do not match the theoretical comparative statics, a theorist can always claim that the measure is "bad" for some other reason—such as complexity, unclear instructions, and so on—and hence the theory is not falsified. However, if a measure lacks construct validity,

Importantly, in our conceptualization, construct validity refers to features of the measure itself, not the data it generates. As such, it can be assessed prior to data collection. Of course, the assessment of construct validity relies on there being testable implications associated with the theoretical construct the measure aims at capturing. For instance, some heavily utilized questions corresponding to fundamental psychological attributes, such as those appearing in the Big Five Personality Test (Roccas et al., 2002), do not correspond to a theoretical model allowing for the assessment of construct validity. Consider, for example, questions asking participants to specify whether they agree with a statements such as, "I complete tasks successfully," or "I like to tidy up," which often constitute part of the Big Five measure of conscientiousness. Comparative statics are not obvious, even across individuals, as those would depend on how words such as "successfully" or "tidy up" are interpreted. Absent a specific theory of conscientiousness, these measures exhibit no testable implications.

In many ways, construct validity is economists' initial "smell test." It is clear to us that a task asking participants to count jelly beans has little to do with risk aversion, whereas a task asking participants to specify certainty equivalents might. The former has no clear association with any feature of risk: any set of responses on any menu of jelly-bean counting tasks would be consistent with any level of risk aversion. On the other hand, certainty equivalents have specific comparative statics dictated by theory.

## 2.2 (Directional) Responsiveness

*(Directional) Responsiveness* is the empirical analogue of construct validity. It reflects whether a measure exhibits empirical patterns in line with the comparative statics on which construct validity rests; whether the resulting data respond to parameter changes as predicted by theory. While the use of the term "Directional" indicates that responses need to be in line with theoretical predictions, we drop this modifier for simplicity.

Within our example of risk attitude measures, responsiveness can be assessed by asking

then failures to support the theory can be more easily blamed on the fact that the measure does not even correctly represent the theory.

participants in an experiment for the certainty equivalent of two (or more) lotteries, each with the same expected payoff, but with a different spread. Assuming utilities are either universally concave or universally convex, if a given participant values a lottery at less than its expected value, then they should also assign lower values to lotteries with the same mean, but greater spread, in their payoffs.

Noise plays an important role in assessing the responsiveness of a measure. It is unrealistic to expect any measure to be fully responsive for all participants, given the potential for participants' misunderstanding, inattention, and other sources of noise. Noisy data lead to attenuated correlations, and may, therefore, weaken comparative statics (Wald, 1940). To mitigate attenuation bias, it suffices to have only two independent measures of the underlying construct. If measurement errors are plausibly independent across these two measures, one can serve as an instrument for the other (Gillen et al., 2019).[3]

The responsiveness of different measures of the same construct can be ranked. A more responsive measure will exhibit more consistent changes across participants in response to similar shifts in the underlying model parameters. This is equivalent to saying that a more responsive measure produces less noisy data. For instance, in the case of risk attitudes, the expected value and spread of payoffs can be characterized for any incentivized measure of which we are aware. One can set up two variants of each of two measures in which the expected value coincides across all four variants, and the variances (and hence difference in variances) are the same across the two measures. The more responsive measure is the one for which responses to the two variants differ by more, and align more reliably with theoretical predictions.

In psychology, what we call "responsiveness" would be considered a key component of construct validity. However, as economists can establish construct validity using theory, we think it is useful to separate the theoretical features of a measure from their empirical implications. For example, a psychologist might argue for the construct validity of a hypothetical measure of aggressiveness by demonstrating that men, who are generally perceived to be

[3]An arguably simpler, albeit less efficient, approach is to average repeat observation of the same measure from the same person.

more aggressive than women, score higher on the proposed measure. Taking this further, a psychologist might also note that individuals convicted of violent crimes score higher on the same measure. A similar approach could be applied to measures of risk aversion, as men are generally believed to be less risk averse than women. One could also compare the risk attitudes of stock traders and those in other professions. Nonetheless, if the only evidence for the validity of a proposed risk aversion measure is that it tends to be higher for women and lower for stock traders, this would provide weak support that the measure truly captures risk aversion. Instead, the observed differences might reflect any number of other characteristics that vary between these groups. In contrast, the approach taken in economics tends to rely on comparative statics derived from a formal theory.

## 2.3 Predictive Validity

*Predictive validity* refers to the extent to which a measure predicts auxiliary variables, behaviors, or outcomes. Predictive validity requires the collection of data relating to both the measure under consideration, and an auxiliary variable or outcome.

The traditional approach to assessing predictive validity is mediated by theoretical models: a measure has high predictive validity if it correlates with other variables that are known or expected to be theoretically related to it. For example, one could evaluate the predictive validity of participation in dangerous sports as a measure of risk attitudes by examining its association with individuals' investment portfolios. Additionally, if two measures assess the same construct and exhibit high predictive validity, they should correlate with each other more than they do with unrelated measures.[4] In our example, we would expect individuals' participation in dangerous sports to correlate more strongly with certainty equivalents of lotteries—presumably capturing an aspect of risk attitudes—than with charitable giving.

A test that depends on the theoretically predicted relationship between measures inher-

[4]This is sometimes referred to as *convergent validity*, see Mata et al. (2018) and Schildberg-Hörisch (2018). It is used in psychology to argue for the construct validity of multiple measures, especially when the underlying psychological constructs do not have a precise definition. As our focus is on evaluating single measures for which construct validity is readily established, it is simpler to see convergent validity as part of predictive validity.

ently relies on the theory underpinning the test. Thus, predictive validation involves the concurrent validation of both the measure and the theory.

Measures with high predictive validity may be more context-dependent than the theories to which they speak. For example, there is a broad theoretical connection between risk attitudes and lottery choices, which transcends language, background, and age. Still, a measure requiring lengthy instructions might predict behavior well among university students but poorly among young children. Theories may be more universally applicable than the measures associated with them.

Of course, predictive validity need not be assessed exclusively within the lab or the survey environment. Associations between a measure and evidence collected outside of the lab could be as valuable. At the same time, evidence outside the lab tends to confound many factors, and does not always offer the most clear-cut approach for validity assessment. For example, suppose a measure of altruism using the dictator game (Forsythe et al., 1994) is not predictive of charitable giving. It might be tempting to see this as evidence that the ultimatum game provides a poor measure of altruism. However, it is possible that altruism is only a secondary motive for charitable giving, an idea that should not be surprising considering the extensive literature exploring the "true" motivations behind charitable gifts; see the review of Andreoni and Payne (2013).

The principle behind predictive validity has been used in the development of new measures. For instance, one might measure risk aversion in a group using an established, responsive measure (with construct validity), then demonstrate that a novel measure of risk aversion, which lacks construct validity—say, a qualitative self-assessment of risk attitudes—produces data that are correlated with the established measure. This process is referred to as "experimental validation" (see Falk et al., 2023). A key requirement is that the two measures must be strongly correlated (once corrected for measurement error); otherwise, any observed correlation between the novel measure and other variables could stem from aspects of the novel measure that do not align with the established measure. Unfortunately, many lab validations produce measures with correlations lower than 0.5. At a correlation of 0.5, it

is equally likely that a correlation between the novel measure and any other variable is due to ancillary variation, rather than variation associated with the construct-valid measure; see Chapman et al. (2025).

As with responsiveness, predictive validity can be evaluated either using measures that generate precise data, or when strategies exist to address measurement noise: for instance, using duplicate elicitations. Naturally, predictive validity has more bite when a measure isolates a single construct. When a measure reflects a combination of constructs, correlations with the measure can be more difficult to interpret. For example, consider a task asking participants to choose one of two lotteries whose realization contributes to the payoff of a different participant. Any choice reflects both risk and social preferences, and an association between this measure and, say, participation in dangerous sports, or charitable giving, is not clear.

The determination of what constitutes sufficient correlation for predictive validity is, in many ways, subjective. However, many correlations perceived as very strong fall between 0.3-0.5. For instance, the correlation between parents' and their children's heights is approximately 0.50 (Wright and Cheetham, 1999); and the correlation between parents' average education level and their children's education ranges from about 0.30 in Denmark to 0.54 in Italy, with most Western countries falling in between (Hertz et al., 2008). Paradoxically, very high correlations may raise suspicion that the measure and the variable it correlates with are capturing the same construct, rather than two associated constructs.[5]

## 2.4 Monetary, Time, and Complexity Costs

The *cost* of a measure may entail the incentives it requires, its complexity, and the time it takes to execute.

Researchers usually face financial and time constraints in experimental settings. A measure that requires substantial incentives, or takes a long time to execute, would necessarily

[5]If the variable being predicted is a behavior, then it is sometimes more efficient to measure the variable directly. This is true even for behaviors and beliefs that some academics suspect people will not readily admit, such as racial prejudice, see Peyton and Huber (2021).

introduce trade-offs in terms of the amount of data that can be collected, on the measure itself or on other measures. More broadly, costly measures may contribute to professional inequalities: only scholars with hefty budgets and teams of assistants can carry out expensive and lengthy data collection efforts.

Efficient utilization of emerging participant pools presents a promising avenue for reducing costs. The advent of online experimental platforms has created new opportunities for obtaining lower-cost data. Nonetheless, some evidence suggests that data from online samples may entail greater noise relative to data from traditional labs (see Snowberg and Yariv, 2021; Fréchette et al., 2022).

For participants, complex or lengthy tasks could lead to confusion and impatience, which can ultimately contribute to noise in the generated data.[6] Furthermore, responses to complex tasks may reflect a mixture of underlying behaviors with heuristics used in responding to a specific elicitation. For instance, when participants specify certainty equivalents of compound lotteries—which are arguably of greater complexity than "simple" lotteries—their responses are strongly associated with certainty equivalents for related ambiguous lotteries (see Halevy, 2007, and Gillen et al., 2019).

We view the refinement of measures to reduce time or complexity costs as part of the evolution of the field and an interesting area of research in itself. For example, Niederle and Vesterlund (2007) developed a measure for assessing a preference for competition. Their experiment had an average runtime of approximately 45 minutes. By using fewer and shorter tasks, Gillen et al. (2019) reduced the average time participants spent on the competition task to around 8 minutes, generating similar results.[7] In the context of belief elicitation, covered in detail in Chapter 3, Wilson and Vespa (2018) suggest a simplified description of the binarized scoring rule offered by Hossain and Okui (2013); see also Danz et al. (2022).

[6] Complexity can be difficult to anticipate. Nonetheless, data can often be used to indicate whether a task is complex. For instance, participants' reliance on detectable response heuristics—for instance, responses that are salient in their choice menu, such as mid-points or end-points of an interval—can suggest increased complexity. See Chapman et al. (2024), Chapman et al. (2025), and our discussion in Section 3.2 for examples.

[7] Qualitative self-reports of competitiveness, such as, "Competition brings the best out of me," are much faster to administer (see Fallucchi et al., 2020). However, this particular question predicts only 26% of the variation in an incentivized competitiveness task, raising concerns about whether predictive validity is due to variation in the underlying task or to other variation in this novel measure.

Efficient utilization of emerging participant pools presents another promising avenue for reducing costs. The advent of online experimental platforms has created new opportunities for obtaining lower-cost data. Nonetheless, some evidence suggests that data from online samples may entail greater noise relative to data from traditional labs (see Snowberg and Yariv, 2021; Fréchette et al., 2022). This handbook provides additional suggestions for reducing the costs of experimental elicitations.

## 2.5 Reliability

The *reliability* of a measure reflects the similarity of repeated uses of a measure on the same object—in the case of the social sciences, a person—within a short period of time: within an experimental session, or over a few days.

Reliability of continuous measures, or discrete measures with a large number of possible values, can be expressed via a Pearson correlation or Spearman rank-order correlation. Much like physical measurements—two length measurements of the same object are expected to coincide up to the smallest unit marked on a ruler—reliability encompasses the classical measurement criteria of precision. Assuming the construct considered remains stable within duplicate executions of the measure, distance between its outputs indicates the precision of, and the noise in, the measure.

We consider reliability to be of secondary importance, as it is generally straightforward to enhance, albeit at some cost. For example, reliability can be mechanically increased by presenting respondents with multiple similar questions and averaging their responses. This practice of quasi-repetition is commonly used in the creation of psychological scales. For example, the Big Five Personality Test for "Openness" contains the following questions:

> Agree or disagree: I have a vivid imagination.
>
> Agree or disagree: I am full of ideas.
>
> Disagree or agree: I do not have a good imagination.
>
> Disagree or agree: I am not interested in abstract ideas.

Disagree or agree: I have difficulty understanding abstract ideas.

While these questions exhibit subtle distinctions, each provides a related measure of "Openness." Averaging the responses yields a more reliable measure. This quasi-repetition strategy is particularly valuable when it is crucial to accurately measure a construct within an individual. For instance, investment funds frequently rely on asset allocation guides, which assess individuals' approach to risk through a series of verbal and quantitative questions, ultimately generating a score that aggregates their responses.[8] As retirement savings affect individuals throughout their lives, it is important to tailor investments to their risk tolerance.

Reliability is not a particularly important concern when one aims to estimate a correlation between a measure and some other variable, nor if one wishes to use a measure as a control. Low reliability can bias results, but two measures of the underlying construct alleviate this bias, as noted in Section 2.2.

The reliability criterion is sometimes applied to classification schemes, where multiple raters classify data that is prone to subjective interpretation. Research assistants, human or artificial, often classify text, images, facial expressions, and so on. *Inter-rater reliability* then refers to the extent to which classifications between different raters are correlated, but is, at least in principle, independent of the quality of classification scheme employed.[9]

## 2.6 Stability

The *stability* of a measure reflects the similarity of repeated uses of a measure on the same object within medium to long time horizons.

In general, a reasonable measure should not be more stable than the underlying construct it aims to measure. In certain fields, such as personality research, scholars particularly value stable measures. If one believes the underlying construct, like personality, is unchanging, an unstable measure would not satisfy our responsiveness criterion—in particular, it would

[8] See, for example, the TIAA Asset Allocation Guide.

[9] We encourage scholars to employ external classifiers when coding data that is susceptible to subjective interpretation. With such data, there is a risk that even the most well-intentioned scholars will subconsciously generate classifications in a motivated fashion.

exhibit changes when theory says it should not. In economics, there is ongoing research aimed at understanding the stability of economic preferences and behaviors.[10] Thus, we believe one should not prioritize stability in a measure at the expense any of the prior criteria. Doing so would hinder the understanding of potential dynamics of economic preferences and behaviors.

The observed stability of a measure can be impacted by its reliability. If a measure generates noisy data at two dates in time, the correlation will be attenuated and may lead to a false impression of low stability. As noted in Section **??**, duplicate measures at each point in time provide a sufficient correction for reliability to allow for stability to be assessed properly.

## 2.7 Summary

Table 1 summarizes some of the important features of the various criteria above. These criteria require different methods of assessment, and need not be aligned with one another. They are also not binary: what seems like a low correlation or low costs for one researcher may seem high for another.

Some of these criteria naturally fall into a partial hierarchy. Construct validity stands as the most fundamental and, in principle, the easiest to assess, as it requires no data collection. Both responsiveness and predictive validity depend on construct validity, with responsiveness serving as its empirical counterpart. Predictive validity demands additional data, and may be more challenging to evaluate, depending on the context. Reliability, while relatively easy to generate through duplicate elicitations, and, thus, arguably less critical when developing a new measure, is useful for assessing both responsiveness and predictive validity. Without reliable responses, correlations with underlying parameters or auxiliary variables may be attenuated. We consider stability the least critical, partly because existing evidence indicates that preferences can be somewhat malleable and influenced by experiences. Monetary, time, and complexity costs are harder to rank within this framework, as they often hinge on the

[10] For example, in the context of risk preferences, Malmendier and Nagel (2011) study how early life experiences influence later-life risk attitudes, while Friedman et al. (2014) discuss stability of risk attitudes in experimental work.

specific circumstances faced by researchers.

Ultimately, researchers will often need to balance trade-offs and apply judgment. We hope our classification provides a language for explicitly articulating these trade-offs and judgment calls. We now describe several examples using these criteria.

Table 1: Summary of Economic Measurement Criteria

| **Measure** | **Data Required for Assessment** | **Model Specific** | **Sample Specific** | **Improvable** |
|---|---|---|---|---|
| Construct Validity | No | Yes | No | No |
| Responsiveness | Yes | Yes | Yes | No |
| Predictive Validity | Yes | Yes | Yes | No |
| Monetary, Time, and Complexity Costs | No | No | Somewhat | Clever designs help |
| Reliability | Yes | No | Yes | Yes, via quasi-repetitions |
| Stability | Yes | No | Yes | No |

# 3 Illustrative Applications

We examine three examples: the desirability of measures commonly deployed in lab versus field experiments, the impact of "clumpy" responses on an experimentalist's evaluation of a measure, and the approach to assessing incentives in economic experiments.

## 3.1 Measures in the Lab and in the Field

The measurement criteria outlined above help refine the ongoing discussion about the relative merits of lab and field experimental approaches.

This discussion has two key facets: the participant population (see, for example, Levitt

and List, 2007, 2009) and the type of measures used (Harrison and List, 2004, see, for example). While measurement criteria offer limited guidance on selecting participant populations, two points are worth noting. First, empirical evidence suggests that lab participants—specifically, students—do not differ significantly from broader populations (see Snowberg and Yariv, 2021, and references there). Second, some field experiments can only be conducted with specialized populations. For example, the groundbreaking study of List (2003) examined the effects of market experience on the endowment effect, using a population of commemorative pin traders. Whether this population is more representative of an economically relevant population than students remains an open question.

Experiments can also be classified by the type of measures they employ: lab-type measures, which involve abstract tasks like giving in the dictator game, or field-type measures, which engage participants in tasks they may encounter in their day-to-day life, such as donating to charity.[11] The mapping between the experimental setting and the types of measures used is not one-to-one: some lab experiments include day-to-day tasks, while some field experiments assess responses to abstract tasks. Fundamentally, the choice between these types of measures reflects a trade-off: lab-type measures prioritize construct validity and responsiveness, whereas field-type measures emphasize predictive validity.

Lab-type measures, designed for controlled environments, typically ensure construct validity by design. Furthermore, some lab-type measures are also examined for responsiveness. For example, in a lab setting, researchers can elicit certainty equivalents for lotteries that are mean preserving spreads of one another, a standard approach in experimental economics (see Roth and Kagel, 1995). In contrast, establishing construct validity in the field is far more challenging, as the necessary comparative statics are rarely observed. Using the certainty equivalent example, one would need to identify a real-world setting where individuals price lotteries with varying characteristics while holding all other factors constant. Without the possibility of testing comparative statics, construct validity is lost.[12]

[11] This distinction parallels Harrison and List's (2004) classification of field experiments on the basis of the measures they use.

[12] Construct validity requires that a measure impose a theoretically testable restriction. Nevertheless, if this restriction cannot be tested in practice, we view the measure as failing the standard of construct validity.

On the other hand, field-type measures are often asserted to have greater predictive validity than lab-type measures.[13] However, this assertion has received limited empirical scrutiny. Notably, a growing body of research suggests that, in fact, behaviors and associations observed in lab settings closely resemble those found in the field and across diverse participant samples (Armantier and Boly, 2013; Alm et al., 2015; Herbst and Mas, 2015).

In particular cases, predictive validity is more critical than construct validity. For instance, when evaluating the potential impact of an educational subsidy, a targeted field study—commonly known as an *impact evaluation*—is often the preferred approach. Such studies typically implement the subsidy for a randomly selected population, allowing researchers to assess its effects in a setting almost identical to the one where the proposed subsidy would actually be implemented (see, for example, Behrman et al., 2005).

In many cases—especially when the goal is to test a theory—construct validity takes precedence. In such instances, lab-type measures are often the most appropriate starting point. For example, to examine the relationship between a particular auction format and bidding behavior, a controlled setting provides a clearer test by eliminating confounding factors, such as unobserved preferences and information. In contrast, studying real-world auctions, where these elements are often unobserved, would complicate empirical analysis.

Between an initial theoretical test and an impact evaluation, the choice of experiment depends on various measurement criteria, with cost being a key consideration. Continuing with the auction example, how should one identify a revenue-maximizing design? Field experiments are typically expensive, and their added predictive validity may not always justify the cost. A more efficient approach might be to first conduct a lab experiment using field-type measures to evaluate multiple designs, narrowing the options to the most promising ones. A subsequent field experiment could then determine which of these finalists maximizes revenue. This approach has been taken in practice. For instance, Goeree and Holt (2005) and Porter and Smith (2006) describe the utilization of lab experiments in the design of FCC auctions.

[13]The assertion is often framed under the umbrella of *external validity*. This term has many definitions, often encompassing sample effects and issues pertaining to predictive validity.

## 3.2 Focal Value Response

It is well known that individuals selecting an alternative from a list or range of choices tend to choose "focal values:" in particular the top, bottom, and middle alternatives (see, for example, Schwarz and Oyserman, 2001).[14] This pattern of choices is not specific to discrete elicitations. For instance, elicitations using "convex time budgets" appear to suffer from similar problems: Andreoni and Sprenger (2012) report that 70% of all choices in their experiments were focal, corner alternatives, and 37% of participants chose a focal value in every decision they faced. In the "risky project" task of Gneezy and Potters (1997), Chapman et al. (2024) find that 60% of responses were to invest 0, 50, or 100% of the investment budget, which they refer to as *focal value response* (FVR).[15]

While some experimentalists view FVR with concern, is it truly problematic? If response patterns, including FVR, genuinely reflect individuals' preferences, then the measure could be considered effective. For example, with reference to the risky project task described above, it might be that 60% of people are distinctly risk averse, moderately risk averse, or not risk averse at all. However, if FVR indicates the use of heuristics or shortcuts in answering questions—potentially akin to Benartzi and Thaler's (2001) 1/n heuristic—this could be seen as a drawback. How can one evaluate these competing possibilities?

Clearly, if one could measure "true" preferences—that is, if one had an accurate measure—then this question would be simple to answer. Unfortunately, as described at the beginning of this chapter, such measures almost never exist. Thus, we turn to the criteria described earlier.

In the absence of a definitive measure of "true" preferences, responsiveness can serve as a useful tool to evaluate whether FVR is problematic. For instance, in the risky project task, one could examine whether the distribution of responses varies when parameters of the problem are adjusted. If a measure displaying FVR shows little to no responsiveness, it

[14] We thank Erin Krupka for suggesting this as a potential application.

[15] Gneezy and Potters's (1997) risky project task provides participants a stock of points (say 100), and allows them to invest however many points they want (between 0 and 100, inclusive) in a project that will pay, for example, three times the amount invested with 35% probability. Any points not invested are kept by the participant. As such, the amount invested gives a measure of risk aversion.

suggests that FVR may be an issue. However, lack of responsiveness is not definitive proof that FVR is the underlying cause; the flaw in the measure could stem from factors unrelated to FVR.

Naturally, if a measure prone to FVR demonstrates responsiveness or predictive validity, it does not necessarily mean that FVR is insignificant. To properly assess the impact of FVR, a comparison with a measure of equal construct validity, but lower FVR incidence, is needed. If this alternative measure proves more responsive or more strongly predictive, it would indicate that FVR does impair the accuracy of the original measure. This would also suggest that the alternative measure is preferable based on responsiveness or predictive validity. However, if the original measure is significantly more cost effective or easier to implement, it may still be the more practical choice.

The other criteria described earlier are of less help in evaluating whether susceptibility to FVR is problematic. In and of itself, this susceptibility does not affect a measure's construct validity, which is assessed using theory. One could attempt to use reliability and stability. However, coarse responses tend to exaggerate correlations, so there is a risk that measures exhibiting FVR would mechanically generate relatively high reliability and stability.[16]

## 3.3 Do Monetary Incentives Matter?

A longstanding debate in experimental economics concerns the necessity of monetary incentives in eliciting preferences from participants. The central question is whether participants need to have payments linked to their choices in order to commit the required time and effort to identify their preferences or express them, or if hypothetical stakes suffice. The discussion often implicitly hinges on the criteria above. For instance, hypothetical incentives typically entail lower costs, both in terms of monetary outlay and in facilitating access to participant pools where monetary incentives may not be feasible. However, cost-effectiveness alone is generally not deemed sufficient to justify the use of hypothetical incentives.

[16]This is, of course, not a given. For example, if individuals randomly choose one of two options out of a large number of possible focal responses, the measure would exhibit less reliability and/or stability.

Using hypothetical incentives maintains construct validity if one assumes participants treat hypothetical and real stakes similarly: the theory that justifies the incentivized version of the measure can still be used to justify its hypothetical counterpart.

Traditional arguments about incentives suggest that hypothetical incentives may diminish responsiveness and predictive validity. Indeed, starting from Camerer and Hogarth (1999), some evidence indicates that lower incentives yield noisier responses. Consequently, the correlations between responses and auxiliary variables are attenuated. Even so, the degree to which this impacts the usefulness of a measure depends on the application. For instance, macroeconomists sometimes use elicited preference distributions to calibrate key parameters (Stango and Zinman, 2020). In such cases, a measure that reflects the distribution of a particular preference—even one that misrepresents individual preferences—may be considered "good enough." Conversely, when the goal is to target interventions or new technologies to individuals with specific preferences (Andreoni et al., 2023), the hypothetical measure must closely predict the incentivized measure at the individual level to be deemed sufficient.

# 4 Choosing Design Parameters

Economic experiments often assess complex behavior, at times involving multiple agents, within a theoretical framework: how agents behave in markets (Smith, 1989), in two-sided matching settings (Echenique et al., 2016, 2024), in election environments (Blais et al., 2016), and so on.

Of the six criteria we spelled out earlier, reliability is arguably the most prominent consideration when designing complex experiments. Indeed, construct validity holds almost by definition, as the actions assessed tend to map directly to a theoretical framework.[17] Moreover, responsiveness and predictive validity are often established through two standard practices. The first is by examining how behavior changes in response to simple manipulations,

[17]This is true even in settings that identify new "behavioral" tendencies—time-inconsistent preferences, altruism, and so on—as they are often assessed by the degree to which behavior deviates from an explicit theory: normative rationality.

essentially testing comparative statics of the underlying model. For example, in a voting setting, experimenters might inspect whether increases in participation costs—a simple design parameter present in many voting experiments—yield decreases in participation (Agranov et al., 2018). The second method of establishing predictive validity examines whether behavior accords with related measures. It does so by comparing results with those of a prior, similar, experiment. Experimental papers building on prior work frequently replicate variants of the original design, with possibly different parameters and interface. This common practice is sometimes referred to as *quasi-replication*.[18] For a more elaborate discussion on issues pertaining to replicability, see Chapter 11.

Finally, time and complexity costs are—similar to construct validity—addressed by standard practices: experimenters pore over instructions, make sure participants complete sessions within a reasonable amount of time, and often include attention or comprehension quizzes. Stability is rarely presumed, and tends to be of secondary concern.

The main challenge in designing complex experiments is that the underlying models that guide the design often provide limited information on how reliable—that is, noisy—participants' responses will be. Careful selection of parameters can assist with designs that are likely to yield economically and statistically significant results.

In this section, we spell out several considerations that may guide the choice of parameters. We believe making those explicit may help assess the quality of experimental designs and the generalizability of the results they produce. As the parameter-selection problem tends to vary with the purpose of the experiment, we also propose a typology of complex experiments that should help experimenters identify which formulation best fits their goals.

Throughout, we consider a setting in which participant $i$'s reward is given by $V_i(\theta, a)$, where $\theta$ is the model parameter, which may be multi-dimensional, and $a$ is the action profile within the group that participant $i$ interacts with. For simplicity, we consider only choices that are static in nature. In particular, all interacting participants make choices simulta-

[18]For instance, Agranov and Tergiman (2014) allow for communication prior to a coalitional bargaining game, quasi-replicating prior treatments carried out by Fréchette et al. (2003), by eliminating communication in one of their treatments; Kübler and Weizsäcker (2004) quasi-replicate the social learning treatments of Anderson and Holt (1997); and so on. For several other examples, see Fréchette et al. (2022).

neously.[19] Further, we assume $a_i^*(\theta)$, the equilibrium—or, in the case of individual decision making, the optimal—choice of participant $i$, is determined uniquely. Otherwise, any test of the underlying model would effectively introduce a confounding coordination problem. The main challenge, then, is which parameter $\theta$ to select, which will depend on the experiment's objective.

## 4.1 Documenting Behavioral (Ir)regularities

When attempting to document a new behavioral regularity, experimenters often try to document the divergence from some benchmark $\overline{a}$: traditionally, normative rationality, but increasingly, more nuanced models.[20] In the context of strategic interaction, experimenters frequently consider behavioral models of interaction as alternatives to classical models of strategic behavior (for a variety of early examples, see Camerer, 2011). In this subsection, we consider symmetric settings with a unique symmetric equilibrium, so we drop the dependence of equilibrium choices on participants' labels.Under our assumptions, the parameter-choice problem is then quite simple: choose

$$\operatorname*{argmax}_{\theta} ||a^*(\theta) - \overline{a}||,$$

in which $||\cdot||$ is a metric on the space of action profiles. Namely, one should choose parameters that maximize the distance from the benchmark when agents act optimally or in equilibrium. In principle, this gives the researcher the best shot at detecting a behavioral regularity, assuming the experimenter's beliefs about how participants respond to changes in $\theta$ is correct.

As an example, consider a social learning experiment à la Anderson and Holt (1997), which aimed to understand whether people "rationally herd" (as in Bikhchandani et al.'s,

[19] Many experiments involve non-trivial heterogeneity and dynamics that allows for learning from or signaling by others. Our discussion below is relevant for such experiments as well. Indeed, the action of each agent $i$, $a_i$ could, in principle, be a mapping from private information to choices, or a contingent dynamic plan chosen at the outset. We maintain this general and simplified notation for clarity's sake.

[20] When using a benchmark of normative rationality, these are often described as anomalies or irregularities. We use the term regularity to focus on the behavior itself, rather than the benchmark against which it is being judged.

1992 model), or conform to others' actions for some other reason. In this experiment/model, individuals start with a prior belief about which of two alternatives is superior. In sequence, each observes her predecessor's choices and a private, conditionally i.i.d. signal before making a choice.[21] If the benchmark action $\overline{a}$ is conformist—say, individuals choose the action corresponding to the majority action chosen before—extremely precise private signals would not be a wise choice according to our criterion. Indeed, Bikhchandani et al. (1992) show that with rational agents and precise signals, *cascades*—constant action choices—commence once two consecutive agents choose the same action. With precise signals, even optimal behavior would lead cascades to form early on, and hence would leave little chance to observe "excess" conformity.

Figure 1 illustrates choices for one-agent problems when the benchmark is $\overline{a} = 0$. In Panel (a), the optimal value under $\theta$ is substantially more distant than that under $\tilde{\theta}$. If one hopes to identify differences from the benchmark behavior, $\theta$ would then be preferable.

## 4.2 Discriminating between Models

Testing between two different models or theories may present conceptual complications. However, these do not generally change the basic logic above. The main complication is that the benchmark $\overline{a}$ is no longer a fixed value, but also changes with the parameter(s) $\theta$. To reflect this, we now write $\overline{a}(\theta)$. The optimization problem is then only slightly modified:

$$\underset{\theta}{\operatorname{argmax}} \, ||a^*(\theta) - \overline{a}(\theta)||.$$

This type of experiment typically emerges when two theories both provide a relatively complete account of behavioral regularities, necessitating novel tests to discriminate between them. For example, Campos-Mercade et al. (2022) aim to distinguish between "classical" prospect theory with a fixed reference point (Kahneman and Tversky, 1979) and a version

[21] One can think of this as a static, symmetric game if agents decide on their full strategies simultaneously, before their order in the sequence is determined.

incorporating an adaptive reference point, à la Kőszegi and Rabin (2006).[22]

In many models, players have different roles, and hence different payoffs. This occurs, for example, in sender-receiver games (Cai and Wang, 2006), network games in which agents hold different network positions (see, Charness et al., 2014, and the discussion in Chapter 7), or in auctions with asymmetric bidders (Avery and Kagel, 1997). Asymmetries can be accommodated by weighting the relative value of observing a difference in behavior in the different roles:

$$\underset{\theta}{\operatorname{argmax}} \sum \alpha_i ||a_i^*(\theta) - \overline{a_i}(\theta)||,$$

in which $\alpha_i > 0$ denotes the weight placed on role $i$. In some instances, for a given role $j$, there are only slight differences in equilibrium play between the two models. Maximizing the objective above will then effectively maximize the difference between equilibrium play across other roles. In such cases, automating the play of agents in role $j$ can be useful: it would likely not limit the insights gained, and potentially reduce noise.[23]

## 4.3 Institutional Design

At times, $\theta$ encompasses an institutional design parameter—say, whether an auction follows a first- or second-price protocol, or whether an election is governed by majority or unanimity rules—in addition to attributes of the environment. In these experiments, an objective is often to maximize some consequences of agents' actions: for instance, revenue in the case of auctions, or welfare in the case of voting rules.

Suppose that $V_i^{SQ}(\theta, a^{SQ}(\theta))$ denotes individual $i$'s utility under the status quo institution, with equilibrium profile $a^{SQ}$, which serves as a benchmark. A researcher who is interested in identifying parameters that maximize welfare relative to the status quo may

[22] Their case is particularly complex as the optimal parameter depends on the individual gain/loss attitudes of participants. Nevertheless, the underlying optimization problem is conceptually the same.

[23] One should utilize automated play with care, as sometimes predicted behavior in a role is not accurate, and deviations are difficult to anticipate. For instance, in classical ultimatum games, one could have expected that, regardless of exchange rates, receivers would accept any amount they are handed. In the lab, experiments repeatedly document receivers conditioning their behavior on what is passed to them; see Cooper and Dutcher (2011).

then consider

$$V(\theta, a) \equiv \sum_i [V_i(\theta, a_i^*(\theta)) - V_i^{SQ}(\theta, a^{SQ}(\theta))]$$

and choose the parameter $\theta^*$ to maximize $V(\theta, a^*(\theta))$. When the parameter $\theta$ pertains only to institutional design parameters, this choice reflects standard mechanism design optimization in a given environment: the optimal auction given bidders' value distributions, or the optimal voting rules given voters' preference distributions.[24] Often, benchmark values of the design parameter that are distant from $\theta^*$ and generate substantially different consequences are also utilized. As an example, Goeree and Zhang (2017) run utilitarian efficient vote buying mechanisms, in addition to a benchmark of simple majority voting.

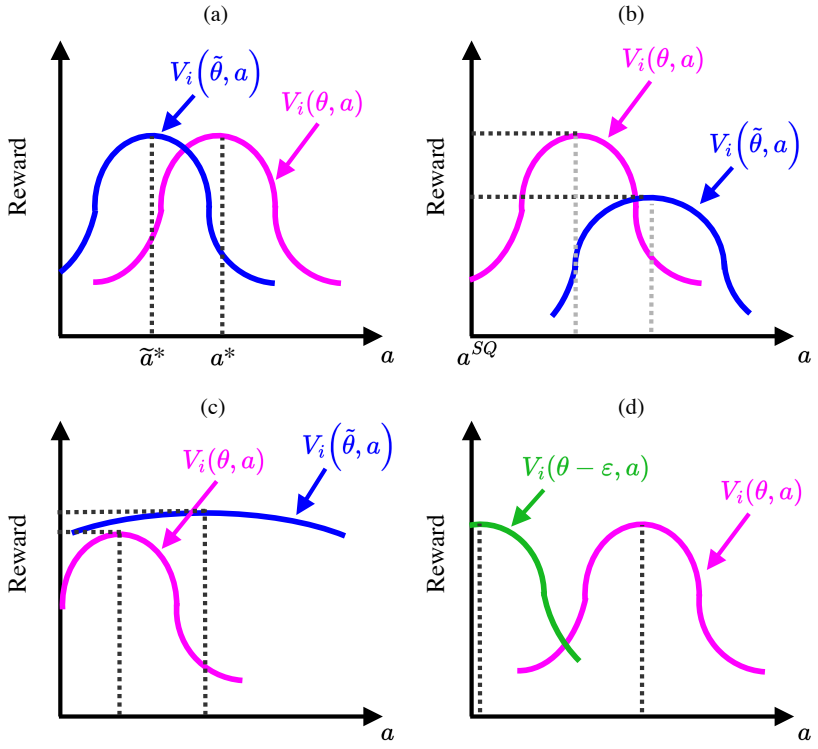
Panel (b) of Figure 1 illustrates this choice, assuming $a_i^{SQ}(\theta) = 0$ for all $i$ and $\theta$, which could serve as a benchmark for action choices. While $\tilde{\theta}$ yields an optimal choice that is more distant than the benchmark relative to $\theta$, the added welfare it generates, assuming individuals choose actions optimally, is lower. If aiming at maximizing the welfare wedge relative to the status quo, implementing $\theta$ would be sensible.

## 4.4 Policy Experiments

Another natural possibility is to choose experimental parameters that echo real-world ones in the setting under consideration. This is particularly appealing when using experiments to speak directly to policy or market design questions. For instance, were one designing new FCC auctions and using multi-unit auction experiments for guidance, mimicking some features of the parameters in the anticipated auctions could be desirable. For a practical example, see Goeree and Holt (2005) and Porter and Smith (2006).

[24] In this case, $\sum_i V_i^{SQ}(\theta, a^{SQ}(\theta))$ would effectively be independent of the design parameter $\theta$ and act as a constant.

Figure 1: Parameter Comparisons




## 4.5 Bringing Reliability Back

In experimental settings, participants often take time to learn how to respond to the environment and how to optimize. *Flat incentives*—where a variety of actions generates very similar rewards—naturally make it more challenging to diagnose and respond to mistakes. This may present a trade-off, as the parameter value that maximizes the distance between, say, the prediction of two models, may also lead to relatively flat incentives. The former parameter will increase discriminatory power, while the latter will decrease it. In particular, an experimenter concerned with reliability would solve the following optimization problem:

$$\underset{\theta}{\operatorname{argmax}} \left|\left| \frac{a_i^*(\theta) - \overline{a_i}(\theta)}{1 + \sigma(a(\theta))} \right|\right|,$$

in which $\sigma(a(\theta))$ is the standard deviation of responses at the parameter value $\theta$, reflecting the reliability of the outcome measure. When the source of variation is the relative steepnesses of incentives, reliability will depend on both the overall incentive level, and $a'(\theta)$, the derivative of $a$ at $\theta$.

The fact that the objective can be clearly written belies the fact that the underlying model will generally provide no information about the outcome measure's reliability, $\sigma(a(\theta))$. While $\sigma(a(\theta))$ is presumably (hopefully?) monotonically decreasing in the strength of incentives, the exact rate of this decline is typically unknown and may exhibit "flat spots" or other irregular patterns. In some cases, it might be feasible to identify prior experiments that utilized a similar outcome measure, and try to calibrate $\sigma(a(\theta))$ for varying levels of incentive steepness.

Panel (c) of Figure 1 depicts the trade-offs that might arise when considering the steepness of incentives. While $\theta$ corresponds to optimal choices that are closer than the benchmark and lower welfare relative to $\tilde{\theta}$, incentives are far steeper around the optimal choice under $\theta$ than under $\tilde{\theta}$.

Another factor that may lead to changes in the reliability of a measure is misperception of experimental parameters. In particular, participants may perceive experimental parameters with some error, even when those parameters are explained in great details during the experiment's instruction phase. If $a^*(\theta)$ is very sensitive to the specification of $\theta$, observing choices different than optimal would be challenging to decipher. Another consideration that may complicate the objective described above is, therefore, robustness to "small" misperceptions.

Panel (d) of Figure 1 illustrates this robustness concern. Even a slight misperception of $\epsilon$ leads to a substantial shift in the optimal action, making it appear noticeably different from the optimal action corresponding to the true $\theta$.

**Avoiding corner solutions** When actions are taken from a finite set—such as probability assessments or allocations of a surplus/budget between 0 and 1—it is generally advisable to avoid parameters that would likely yield outcomes near the "corners" of the range. For

example, when dealing with actions within the $[0, 1]$ interval, parameters $\theta$ should be chosen carefully to avoid cases where $a^*(\theta) \approx 0$ or $a^*(\theta) \approx 1$. This caution is due to the potential for a *Compression Effect* due to truncation. Specifically, consider a setting where a participant reports

$$\tilde{a}^*(\theta) = \min\{\max\{a^*(\theta + \epsilon), 0\}, 1\},$$

with $\epsilon$ capturing non-trivial random noise. If $a^*(\theta)$ is near 0 or 1, average reports would be compressed toward the center of the interval $(0, 1)$ and appear to the researcher as distinct from the optimal choice (see Enke and Graeber, 2023). The induced truncation can impact measurement reliability, potentially increasing noise in choices near the range's boundaries.

# 5 Discussion

In this chapter, we explored the essential criteria for designing and evaluating new experimental measures. We also delved into the art of parameter selection in complex experiments, highlighting the importance of thoughtful design choices. By considering the nuances of each criterion, researchers can fine-tune their experiments to yield more accurate and meaningful insights. Whether it is balancing the trade-offs between cost and reliability, or adjusting parameters to avoid the pitfalls of "clumpy" responses, the guidance provided here is intended to equip scholars with the tools to make informed, strategic decisions.

It is important to remember that the path to discovery is as much about how you ask questions as it is about the questions themselves. By applying the principles discussed in this chapter, we hope you will be better prepared to design experiments that push the boundaries of what we can learn about human behavior and economic decision-making.

In the chapters that follow, you will find a diverse array of methodologies and case studies, each contributing to a richer understanding of experimental economics. We invite you to explore these perspectives, integrate these tools into your own research, and continue the conversation about how best to design measures and experiments that reflect behavior and the complexities of the real world.

# References


**Agranov, Marina and Chloe Tergiman**, "Communication in Multilateral Bargaining," *Journal of Public Economics*, 2014, *118*, 75–85.

_ **, Jacob Goeree, Julian Romero, and Leeat Yariv**, "What Makes Voters Turn Out: The Effects of Polls and Beliefs," *Journal of the European Economic Association*, 2018, *16* (3), 825–856.

**Alm, James, Kim Bloomquist, and Michael McKee**, "On the External Validity of Laboratory Tax Compliance Experiments," *Economic Inquiry*, 2015, *53* (2), 1170–1186.

**Anderson, Lisa and Charles Holt**, "Information Cascades in the Laboratory," *The American Economic Review*, 1997, pp. 847–862.

**Andreoni, James and A.Ãbigail Payne**, "Charitable Giving," in Alan Auerbach, Raj Chetty, Martin Feldstein, and Emmanuel Saez, eds., *Handbook of Public Economics*, Vol. 5, Elsevier, 2013, pp. 1–50.

_ **and Charles Sprenger**, "Estimating Time Preferences from Convex Budgets," *The American Economic Review*, 2012, *102* (7), 3333–3356.

_ **, Michael Callen, Karrar Hussain, Muhammad Yasir Khan, and Charles Sprenger**, "Using Preference Estimates to Customize Incentives: An Application to Polio Vaccination Drives in Pakistan," *Journal of the European Economic Association*, 2023, *21* (4), 1428–1477.

**Armantier, Olivier and Amadou Boly**, "Comparing Corruption in the Laboratory and in the Field in Burkina Faso and in Canada," *Economic Journal*, 2013, *123* (573), 1168–1187.

**Avery, Christopher and John Kagel**, "Second-Price Auctions with Asymmetric Payoffs: An Experimental Investigation," *Journal of Economics & Management Strategy*, 1997, *6* (3), 573–603.

**Behrman, Jere, Piyali Sengupta, and Petra Todd**, "Progressing through PROGRESA: An Impact Assessment of a School Subsidy Experiment in Rural Mexico," *Economic Development and Cultural Change*, 2005, *54* (1), 237–275.

**Benartzi, Shlomo and Richard Thaler**, "Naive Diversification Strategies in Defined Contribution Saving Plans," *American Economic Review*, 2001, *91* (1), 79–98.

**Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch**, "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades," *Journal of Political Economy*, 1992, *100* (5), 992–1026.

**Blais, André, Jean-François Laslier, and Karine Van der Straeten**, *Voting Experiments*, Springer, 2016.

**Cai, Hongbin and Joseph Tao-Yi Wang**, "Overcommunication in Strategic Information Transmission Games," *Games and Economic Behavior*, 2006, *56* (1), 7–36.

**Camerer, Colin**, *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton University Press, 2011.

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

**Campos-Mercade, Pol, Lorenz Goette, Thomas Graeber, Alexandre Kellogg, and Charles Sprenger**, "Heterogeneity of Gain-loss Attitudes and Expectations-based Reference Points," 2022. Mimeo.

**Chapman, Jonathan, Erik Snowberg, Stephanie Wang, and Colin Camerer**, "Dynamically Optimized Sequential Experimentation (DOSE) for Estimating Economic Preference Parameters," 2024. Mimeo.

_ **, Pietro Ortoleva, Erik Snowberg, Leeat Yariv, and Colin Camerer**, "Reassessing Qualitative Self-Assessments and Experimental Validation," *Mimeo*, 2025.

**Charness, Gary, Francesco Feri, Miguel Meléndez-Jiménez, and Matthias Sutter**, "Experimental Games on Networks: Underpinnings of Behavior and Equilibrium Selection," *Econometrica*, 2014, *82* (5), 1615–1670.

**Cooper, David and Glenn Dutcher**, "The Dynamics of Responder Behavior in Ultimatum Games: A Meta-Study," *Experimental Economics*, 2011, *14*, 519–546.

**Danz, David, Lise Vesterlund, and Alistair Wilson**, "Belief Elicitation and Behavioral Incentive Compatibility," *American Economic Review*, 2022, *112* (9), 2851–2883.

**Echenique, Federico, Alejandro Robinson-Cortés, and Leeat Yariv**, "An experimental study of decentralized matching," *Working paper*, 2024.

_ **, Alistair Wilson, and Leeat Yariv**, "Clearinghouses for Two-sided Matching: An Experimental Study," *Quantitative Economics*, 2016, *7* (2), 449–482.

**Engel, Christoph**, "Dictator Games: A Meta Study," *Experimental economics*, 2011, *14*, 583–610.

**Enke, Benjamin and Thomas Graeber**, "Cognitive Uncertainty," *The Quarterly Journal of Economics*, 2023, *138* (4), 2021–2067.

**Falk, Armin, Anke Becker, Thomas Dohmen, David Huffman, and Uwe Sunde**, "The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences," *Management Science*, 2023, *69* (4), 1935–1950.

**Fallucchi, Francesco, Daniele Nosenzo, and Ernesto Reuben**, "Measuring Preferences for Competition with Experimentally-Validated Survey Questions," *Journal of Economic Behavior & Organization*, 2020, *178*, 402–423.

**Forsythe, Robert, Joel Horowitz, Nathan Savin, and Martin Sefton**, "Fairness in Simple Bargaining Experiments," *Games and Economic behavior*, 1994, *6* (3), 347–369.

**Fréchette, Guillaume, John Kagel, and Steven Lehrer**, "Bargaining in Legislatures: An Experimental Investigation of Open versus Closed Amendment Rules," *American Political Science Review*, 2003, *97* (2), 221–232.

_ **, Kim Sarnoff, and Leeat Yariv**, "Experimental Economics: Past and Future," *Annual Review of Economics*, 2022, *14* (1), 777–794.

**Friedman, Daniel, R.M̃ark Isaac, Duncan James, and Shyam Sunder**, *Risky Curves: On the Empirical Failure of Expected Utility*, Routledge, 2014.

**Gillen, Ben, Erik Snowberg, and Leeat Yariv**, "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study," *Journal of Political Economy*, 2019, *127* (4), 1826–1863.

**Gneezy, Uri and Jan Potters**, "An Experiment on Risk Taking and Evaluation Periods," *The Quarterly Journal of Economics*, 1997, *112* (2), 631–645.

**Goeree, Jacob and Charles Holt**, "Comparing the FCC's Combinatorial and Non-combinatorial Simultaneous Multiple Round Auctions: Experimental Design Report," 2005.

_ **and Jingjing Zhang**, "One Man, one Bid," *Games and Economic Behavior*, 2017, *101*, 151–171.

**Güth, Werner, Rolf Schmittberger, and Bernd Schwarze**, "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior & Organization*, 1982, *3* (4), 367–388.

**Halevy, Yoram**, "Ellsberg Revisited: An Experimental Study," *Econometrica*, 2007, *75* (2), 503–536.

**Harrison, Glenn and John List**, "Field Experiments," *Journal of Economic Literature*, 2004, *42* (4), 1009–1055.

**Herbst, Daniel and Alexandre Mas**, "Peer Effects on Worker Output in the Laboratory Generalize to the Field," *Science*, 2015, *350* (6260), 545–549.

**Hertz, Tom, Tamara Jayasundera, Patrizio Piraino, Sibel Selcuk, Nicole Smith, and Alina Verashchagina**, "The Inheritance of Educational Inequality: International Comparisons and Fifty-year Trends," *The BE Journal of Economic Analysis & Policy*, 2008, *7* (2).

**Hossain, Tanjim and Ryo Okui**, "The Binarized Scoring Rule," *Review of Economic Studies*, 2013, *80* (3), 984–1001.

**Johnson, Noel and Alexandra Mislin**, "Trust Games: A Meta-analysis," *Journal of Economic Psychology*, 2011, *32* (5), 865–889.

**Kahneman, Daniel and Amos Tversky**, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 1979, *47* (2), 263–291.

**Kőszegi, Botond and Matthew Rabin**, "A Model of Reference-Dependent Preferences," *The Quarterly Journal of Economics*, 2006, *121* (4), 1133–1165.

**Kübler, Dorothea and Georg Weizsäcker**, "Limited Depth of Reasoning and Failure of Cascade Formation in the Laboratory," *The Review of Economic Studies*, 2004, *71* (2), 425–441.

**Levitt, Steven and John List**, "What do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?," *Journal of Economic Perspectives*, 2007, *21* (2), 153–174.

_ **and** _ , "Field Experiments in Economics: The Past, the Present, and the Future," *European Economic Review*, 2009, *53* (1), 1–18.

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

**Malmendier, Ulrike and Stefan Nagel**, "Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?," *The Quarterly Journal of Economics*, 2011, *126* (1), 373–416.

**Mata, Rui, Renato Frey, David Richter, Jürgen Schupp, and Ralph Hertwig**, "Risk Preference: A View from Psychology," *Journal of Economic Perspectives*, 2018, *32* (2), 155–172.

**Niederle, Muriel and Lise Vesterlund**, "Do Women Shy Away from Competition? Do Men Compete too Much?," *The quarterly journal of economics*, 2007, *122* (3), 1067–1101.

**Peyton, Kyle and Gregory Huber**, "Racial Resentment, Prejudice, and Discrimination," *The Journal of Politics*, 2021, *83* (4), 1829–1836.

**Porter, David and Vernon Smith**, "FCC License Auction Design: A 12-year Experiment," *Journal of Econics & Policy*, 2006, *3*, 63.

**Roccas, Sonia, Lilach Sagiv, Shalom Schwartz, and Ariel Knafo**, "The Big Five Personality Factors and Personal Values," *Personality and Social Psychology Bulletin*, 2002, *28* (6), 789–801.

**Roth, Alvin and John Kagel**, *The Handbook of Experimental Economics*, Vol. 1, Princeton university Press, 1995.

**Schildberg-Hörisch, Hannah**, "Are risk preferences stable?," *Journal of Economic Perspectives*, 2018, *32* (2), 135–154.

**Schwarz, Norbert and Daphna Oyserman**, "Asking Questions about Behavior: Cognition, Communication, and Questionnaire Construction," *The American Journal of Evaluation*, 2001, *22* (2), 127–160.

**Smith, Vernon**, "Experimental Methods in Economics," in "Allocation, Information and Markets," Springer, 1989, pp. 94–111.

**Snowberg, Erik and Leeat Yariv**, "Testing the Waters: Behavior across Participant Pools," *American Economic Review*, 2021, *111* (2), 687–719.

**Stango, Victor and Jonathan Zinman**, "Behavioral Biases are Temporally Stable," 2020. NBER Working Paper #27,860.

**Tversky, Amos and Daniel Kahneman**, "Judgment under Uncertainty: Heuristics and Biases," *Science*, 1974, *185* (4157), 1124–1131.

**Wald, Abraham**, "The Fitting of Straight Lines if both Variables are Subject to Error," *The Annals of Mathematical Statistics*, 1940, *11* (3), 284–300.

**Wilson, Alistair and Emanuel Vespa**, "Paired-uniform Scoring: Implementing a Binarized Scoring Rule with Non-mathematical Language," 2018. Mimeo.

**Wright, Charlotte and Tim Cheetham**, "The Strengths and Limitations of Parental Heights as a Predictor of Attained Height," *Archives of Disease in Childhood*, 1999, *81* (3), 257–260.