Chapter 14
Experimental Data and Econometrics

James R. Bland, Anna Conte, Peter G. Moffatt

Abstract The chapter charts the evolution of the range of econometric techniques that have been applied by Experimental Economists. We attempt to identify the origin of each technique covered in the chapter. Some techniques originated long ago in the statistics or econometrics literatures, and have been borrowed by Experimental Economists as the needs have arisen. Others have originated more recently, in direct response to the needs of Experimental Economists. The most commonly used technique has been treatment testing, and a good deal of attention is paid to studies using this approach, which includes standard parametric and non-parametric treatment tests, tests comparing entire distributions, and bootstrap tests. Closely related to treatment testing is power analysis, and we explain the dramatic increase in the importance of power analysis in the Experimental Economics literature in recent years. The decision over which econometric technique is appropriate for a given application is often determined by the data type. We therefore distinguish the data types arising in Experimental Economics (binary, ordinal, interval, etc.), and identify appropriate techniques, with key examples from the literature. We then consider the recent growth in popularity of fully structural models, with particular attention paid to studies that estimate risk aversion parameters while assuming between-subject heterogeneity in risk aversion. In these contexts, we compare and contrast the method of Maximum Simulated Likelihood (MSL) and Bayesian Hierarchical Models as estimation methods that allow for continuous heterogeneity in preference parameters.

James R. Bland
The University of Toledo, Toledo, OH, USA, e-mail: James.Bland@UToledo.edu

Anna Conte
Sapienza University of Rome, Rome, Italy, e-mail: anna.conte@uniroma1.it

Peter G. Moffatt ✉
University of East Anglia, Norwich, United Kingdom, e-mail: p.moffatt@uea.ac.uk

14.1 Introduction

This chapter attempts to chart the evolution of econometric methods that have been applied to the analysis of data from Economic Experiments. This area of research is sometimes referred to as 'Experimetrics'. A textbook bearing this title was written by Moffatt (2015), followed by similarly titled research monograph (Moffatt, 2021) and two encyclopaedia entries (Moffatt, 2025, 2026).

The earliest recorded analysis of data from an economic experiment was Bernoulli (1738), testing the St. Petersburg paradox. On this basis, Experimetrics could claim to be the oldest area of Econometrics, since this was at least 80 years before the earliest use of regression analysis. However, it must also be said that the rise of Experimetrics has taken place largely since around 1980, lagging the rise of Experimental Economics by around two decades.

Prior to Experimental Economics entering the mainstream, there was a widely-held view that Economics was inequivocally a non-experimental Science. Even as late as 1979, this view was articulated in the hugely popular textbook, *An Introduction to Positive Economics* (Lipsey, 1979), which includes the following:

> *"In some fields, the scientist is able to generate observations that will provide evidence for or against any hypothesis that he wishes to test. Experimental Sciences, such as chemistry and some branches of psychology, have an advantage because it is possible to produce relevant evidence through controlled laboratory experiments. Other sciences, such as astronomy and economics, cannot do this. They must wait for time to throw up observations that may be used to test hypotheses."*[1]

That this view was still prevalent in 1979 is evidence of the sea-change in perspective that has occurred in the last half-century. We no longer need to 'wait for time to throw up observations'. We can obtain as many observations as needed from the lab, and in doing so we enjoy the huge advantage of experimental control. As a result, we can, to a certain extent, stop worrying about some of the econometric problems that have traditionally concerned econometricians, such as sample selection bias, measurement error, and endogeneity.

For the entire period over which Experimental Economics has been practised, a misperception held by many practitioners has been that *all* econometric problems can be side-stepped when data from a controlled experiment is analysed. This misperception derives from the idea that, assuming the treatment effect of interest has been successfully isolated, the only data analysis that is required is the comparison of two samples using a simple treatment test. This view is, of course, far too simplistic. There are many ways of designing an experiment and conducting the required test, and econometric considerations are central to the choices made. And in some types of experiment, advanced estimation methods are required to deal with an array of issues including the structure of dependence in the data resulting from the chosen experimental design, the stochastic nature of decisions, the presence of different types in the population, the presence of between-subject heterogeneity in preference parameters, and the role of learning.

[1] Lipsey (1979), p.8.

We will identify two broad types of Economic Experiment: theory-testing experiments, and 'measurement' experiments. For the first type of experiment, the designs are relatively complex and the required econometric techniques relatively simple (e.g., treatment testing). For the second type of experiment, the designs are relatively simple (e.g., lottery choice or allocation decisions) but, for reasons that will become clear, the required econometric techniques are much more elaborate. For this reason, we pay more attention to the second type of experiment in the chapter.

It must be said that the majority of econometric methods applied in the analysis of Experimental data are methods that previously existed in the econometrics and/or statistics literatures, and have been recently adopted by experimentalists. Only a small number of techniques have been developed exclusively for particular Experimental applications, one notable example being Quantal Response Equilibrium (McKelvey & Palfrey, 1995).

14.2 Experimental Design

Experimental Design is a well-developed area of mainstream statistics. A very early contributor was Fisher (1926) who proposed a methodology for designing experiments in agriculture. There followed a vast literature on optimal design applying to both linear models (Silvey, 1980) and non-linear models (Atkinson, 1996). This literature was mainly concerned with the choice of design matrix, that is, the choice of values taken by the independent variables. The concept of a D-optimal design—a design matrix which maximises the determinant of the sample information matrix—was prominent in this literature.

In the context of Experimental Economics, deriving a D-optimal design requires designing a set of tasks that maximise the amount of statistical information resulting from the data. This has been attempted in the context of risky choice, by Moffatt (2007) and Bland (2025a). This application is optimal design theory has proved both interesting and challenging. Because risky choice tasks typically take the form of a choice between two lotteries, the usual desire to have design points as far as possible from each other, occupying the 'corners of the design space', is countered by the desire to have points giving rise to perfect indifference—the requirement of 'utility balance' (Huber & Zwerina, 1996).

There has been some interest in the design of 'interactive experiments', in which subjects' choices are continuously monitored, and used in constructing a design which is locally optimal for the next stage of the experiment (Chaudhuri & Mykland, 1995). An obvious problem with such an approach is that it violates incentive compatibility, if subjects have the opportunity to provide deliberately false responses in order to guide the sequential design towards more favourable tasks, although see Johnson et al. (2021) who develop an approach to designing interactive experiments that avoids this problem.

In the context of Economic Experiments, there are many other aspects to experimental design, relating to the methods of incentivising subjects, and the ways in which subject-groups are chosen in interactive experiments.

In some areas of experimental economics, the focus is on a test of an economic theory, rather than the estimation of 'home-grown' preference parameters of the experimental subjects. The most common types of experiment falling into this category are market experiments (Smith, 1962) and auction experiments (Kagel & Levin, 1986; Ham, Kagel & Lehrer, 2005). In these settings, Induced Value Methodology (Smith, 1962) has been routinely used. This was a major breakthrough for Experimental Economics and helped establish the field as a legitimate scientific research method. It enabled researchers to study markets in a laboratory-controlled setting by using monetary rewards to induce genuine buying and selling behaviour, regardless of individuals' personal preferences. Using this method, researchers were able to test classical price theory in markets and show that markets often converge towards the equilibrium even under conditions of limited information. Smith's contribution brought experimental economics from being a niche to a rigorous and reliable research discipline.

Experimental research often uses the Partner vs. Stranger paradigm to study cooperation, reciprocity, and reputation (see Andreoni, 1988, who formally defined the two protocols). In a partner design, participants interact with the same individual across repeated rounds, encouraging long-term cooperation and reputation building. In contrast, a stranger design randomly rematches participants in each round, creating one-shot interactions that usually reduce cooperation over time. The choice between these two designs depend on the specific research objective and the type of behaviour being analysed. From the econometric perspective, the Partner design is preferred if the focus is on cooperation, but the Partner design results in data with strong forms of dependence that need to be allowed for in estimation and testing.

14.3 Treatment Testing

The most common approach within Experimetrics is the treatment testing approach. In certain types of experiment, there are compelling reasons for considering this to be the natural approach. Provided that subjects are assigned randomly to treatments, and that the experimental environment is such that all influences on behaviour other than the treatment of interest are held fixed, then it is fair to say that there is no need to address many of the types of problem in which econometricians have traditionally been interested, such as sample selection bias, measurement error, and endogeneity.

There are many different approaches to treatment testing. Siegel and Castellan (1988) provided a fairly complete practical guide to these tests, and the fact that this text remains the standard reference after almost four decades provides evidence that the standard methodology has evolved slowly. The first decision to be made is whether a between-subject or within-subject test is required. Between-subject designs assign different participants to different conditions randomly; while within-subject

designs expose participants to all conditions, enabling researchers to control for individual differences (see Charness, Gneezy & Kuhn, 2012, for pros and cons of these experimental designs). Although within-subject tests are more powerful, Experimental Economists have typically favoured between-subject tests. Their main concern with within-subject tests is the distortion brought about by 'order effects' (Holt & Laury, 2005).

Another key decision is between parametric and non-parametric tests. Fully parametric tests such as the t-test rely on strong distributional assumptions, most importantly normality of the decision variable, and in situations in which normality does not hold, they require large sample sizes. In many experimental settings, decision variables are found to be highly non-normal, and for this reason, many experimental researchers have expressed a preference for non-parametric alternatives such as the Mann-Whitney test, which uses ordinal and not cardinal information, and essentially amounts to a test of equality of medians. An alternative approach which does not rely on distributional assumptions but which makes full use of the cardinal information in the data is the bootstrap technique (Ellingsen & Johannesson, 2004). A similar approach which has received attention recently is that of permutation testing (Holt & Sullivan, 2023), the most simple example of which is Fisher's Exact test used for discrete outcomes.

Tests comparing variances have been found to be useful in some settings, for example when one treatment places subjects in a less familiar situation than the other (Moffatt, 2019). In a situation in which it is unclear which functional of the distribution (e.g., mean, median, variance) is impacted by the treatment, a test of the equality of entire distributions may be preferred, the most popular being the Kolmogorov-Smirnov test and the Epps-Singleton test (Forsythe, Horowitz, Savin & Sefton, 1994; Eckel & Grossman, 2000).

Within-subject tests for discrete outcomes have been very important in testing particular types of treatment, for example, testing Allais Paradox (Allais, 1953) and testing the Preference Reversal phenomenon (Grether & Plott, 1979). The most popular test for these purposes has been the McNemar test (McNemar, 1947). An alternative test appropriate for the same type of problem is the Conlisk test (Conlisk, 1989) and this is one example of a statistical test that originated within the Experimental Economics literature.

In many settings, it has become routine to perform treatment tests by testing the significance of a 'treatment dummy' in a regression. Obvious reasons for doing this are that it enables the simultaneous testing of more than one treatment, and it enables the researcher to control for other exogenous factors such as subject-demographics. Another considerable advantage of the regression approach is that it allows the researcher to adjust for dependence between observations. Dependence usually takes the form of clustering, either at the level of the individual subject (as with paired data), or at the level of the group of subjects that are interacting with each other, or at the level of the experimental session (Fréchette, 2012). As emphasized by Moulton (1986), default OLS standard errors that ignore such clustering can greatly underestimate the true OLS standard errors, and can therefore give rise to erroneous detection of non-existent treatment effects.

Of course, it is possible to go further by building the dependence structure into the estimation process. This leads to models such as random effects and multi-level models. Multi-level models are particularly useful because they are flexible enough to allow between-subject heterogeneity in both the intercept and the slope. That is, they allow both the expected response and the treatment effect to vary between subjects. Moreover, the presence of these different types of heterogeneity can be formally tested for. Although this approach has been used by some Experimental Economists (Gächter & Renner, 2010), it has recently been suggested convincingly by Oshchepkov and Shirokanova (2022) that their usefulness dictates that they should be used more often.

14.4 The Rise and Rise of Power Analysis

Closely related to treatment testing is power analysis (Cohen, 2013), which is becoming increasingly important in experimental studies. As evidence of the seriousness with which power analysis is now being taken, in the Editor's Preface of the very first issue (July, 2015) of the *Journal of the Economic Science Association*, the message is very clear:

> "*A necessary (but not sufficient) condition for publishing a replication study or null result will be the presentation of power calculations.*"

Broadly, power analysis can be used for two different purposes: first, to determine the sample size that is required to attain a given, pre-determined, level of power; second, to compute the power of a test that has already been conducted. In both cases, a power of at least 0.8 has become accepted the benchmark for adequacy, although Zhang and Ortmann (2013) have presented evidence that this target has rarely been attained in practice.

For most simple tests, power calculations are simple, and routines for computing power are available in popular software packages. In more complicated settings, for example, treatment tests being conducted in the framework of panel models, multi-level models or structural models, Monte Carlo simulations are useful in performing power calculations. One particularly useful application of Monte Carlo methods in this context is in assessing the relative benefits, in terms of power, of increasing the number of subjects versus increasing the number of tasks (Moffatt, 2021). Another design feature that determines power is the size of the monetary incentives. To take account of this would require incorporating an experimental budget constraint into the power calculation (Alekseev, 2025).

The concept of power plays a key role in the ongoing scientific quality debate (Camerer et al., 2016). One reason for this is that if studies tend to be under-powered, the studies that reach publication will tend to be those with fortuitously high effect sizes. When these high effect sizes are used in power calculations, they will give rise to further under-powered studies (Zhang & Ortmann, 2013).

14.5 Applications of Binary, Ordered, Interval and Censored Models

Many different types of data have arisen in Experimental Economics. The decision variable may be discrete or continuous. A discrete variable may be binary, for example the choice between a risky and safe lottery (Hey & Orme, 1994). Or it may be categorical, for example the choice between a number of different allocations (Engelmann & Strobel, 2004). Or it may be ordinal, for example, when the subject has been asked to reveal their emotions when participating in experimental games (Bosman & Van Winden, 2002). A continuous variable arises when the subject has been asked how much to contribute to a public fund (Bardsley & Moffatt, 2007), or asked for their willingness to pay for an object or lottery (Isoni, Loomes & Sugden, 2011). Even continuous responses contain elements of discreteness, in the form of censoring, lumpiness, and focal points. Methods for allowing for these data features in estimation are covered in detail by Moffatt (2015).

Zero observations are a prominent feature in data from many types of experiment. Sometimes zero observations have a special interpretation. For example, in dictator games, zero transfers are interpreted as 'selfish' behaviour; in public goods games, zero contributions are interpreted as 'free-riding'. In these settings, the tobit model (Tobin, 1958) may be considered the obvious approach. But hurdle models, introduced by Cragg (1971) are often considered more suitable since they allow for 'zero-types' as well as censored zeros. Some researchers have estimated 2-part models (Duan, Manning, Morris & Newhouse, 1983) and presented the results as hurdle-model estimates. This is incorrect because the 2-part model estimates the two hurdles separately while the hurdle model is a full-information model that estimates both hurdles simultaneously.

Since experimental data often consists of a sample of subjects engaging in a sequence of tasks, the panel hurdle model (Dong, Chung & Kaiser, 2004) is usually appropriate. The usefulness of the hurdle approach has been demonstrated by Engel and Moffatt (2012), who apply the model to test for the 'house money effect' using data from a public goods experiment previously conducted by Clark (2002). Clark himself did not find a significant house money effect. Yet this non- result might follow from the fact that he used unconditional non-parametric tests with low statistical power. Engel and Moffatt (2012) apply the panel hurdle model, and find a significant house money effect, but only in the first hurdle. This result was considered particularly interesting because it implied that house money does not simply impact on the contributions of subjects, but it increases the probability of being a free-rider. If the presence of house money has the effect of moving subject between types, this has important implications for the 'external validity' debate (Bardsley et al., 2009) since, obviously in real-world decisions, only 'own money' is present.

In settings where subjects are required to choose between more than two alternatives, discrete choice models are appropriate. One prominent example is Engelmann and Strobel (2004) who analyse choice data on hypothetical income-allocations in order to identify the features of an allocation that are most desirable to the representative

individual. Some researchers have applied discrete choice models in situations in which outcomes can be ordered, or even measured quantitatively. We consider this to be inappropriate. See, for example, Conte and Moffatt (2014)'s critique of Cappelen, Hole, Sørensen and Tungodden (2007).

14.6 Finite Mixture Models (FMM)

Finite Mixture Models (FMM) is a powerful way of dealing with between-subject heterogeneity. The central idea is that the population of subjects is not treated as homogeneous, but as a mixture of a finite number of latent types. Each type is associated with its own behavioural rule. The researcher does not observe a subject's type; rather, type membership is inferred probabilistically from the subject's observed choices. In this sense, FMMs provide a disciplined way for representing such heterogenity without assigning subject to types *ex ante*.

The idea of mixture distributions has a long history in econometrics. The problem of mixture decomposition, that is, the identification of the underlying components of an observed distribution, can be traced back at least as far as Pearson (1894). The earliest appearance of a FMM in Experimetrics was El-Gamal and Grether (1995), and a textbook devoted to the technique was provided by McLachlan and Peel (2000). The appeal of FMMs in Experimetrics is immediate. Experimental data often displays much more heterogeneity than can plausibly be captured by a single representative-agent model: in choice under risk or uncertainty, some subjects behave as Expected Utility maximisers, while others' behaviour is better described by non-Expected Utility models; in social dilemmas, some subjects free-ride, while others cooperate; in strategic interactions, subjects differ in the level of sophistication of their decisions.

It is useful to distinguish between two broad types of finite mixture model in experimetrics. The first is the latent class approach. Here, all subjects are assumed to follow the same behavioural model, but the population is divided into a finite number of classes with different parameter values. That is, subjects respond to the same incentives, but do so in different ways. For example, different classes may have different degrees of risk aversion, different propensities to cooperate, or different levels of strategic sophistication. The classes are not usually given a strong behavioural interpretation before estimation; their role is to approximate unobserved heterogeneity in a parsimonious way. The approach is especially useful when the researcher believes that subjects differ quantitatively rather than qualitatively. The econometrician estimates the parameter values for each type and the mixing probabilities associated with each type. Examples of this kind of mixture model can be found in El-Gamal and Grether (1995) and, more recently, in Bruhin, Fehr and Schunk (2019).

A second type of finite mixture models is more explicitly behavioural. In this case, the types are not merely statistical classes. They are defined by distinct decision rules or preference functionals. The researcher specifies a finite menu of behaviourally

meaningful rules, and the mixture model is used to estimate the proportion of the population who appear to follow each rule. As highlighted by the examples discussed above, this second approach directly puts into practice the core objective of experimental design by allowing different structural models of choice to coexist. For example, in the analysis of choice under risk, Harrison and Rutström (2009) and Conte, Hey and Moffatt (2011), use this framework to allow for both Expected Utility and non-Expected Utility specifications, with the latter combining rule-level qualitative heterogeneity with parameter-level quantitative heterogeneity within each type. Bardsley and Moffatt (2007) and Conte, Levati and Montinari (2019) apply this type of FMM to data from public goods experiments, and identify four types: reciprocator; strategic contributor; altruist; free-rider. Cappelen et al. (2007) and Conte and Moffatt (2014) apply similar models to data from a fairness experiment, and identify 3 types: libertarian; egalitarian; liberal-egalitarian. In strategic-choice experiments, the approach has been used to uncover behavioural strategies or levels of reasoning (El-Gamal & Grether, 1995; Bosch-Domènech, Montalvo, Nagel & Satorra, 2010).

This second type of FMM is also used prevalently in the analysis of direct-response indefinitely-repeated game experiments, where the types are different strategies players may be using in the game. Here, strategies are not observed, but rather we observe the *actions* players choose, which we can think about as the result of two strategies playing against each other. 'Strategy Frequency Estimation' (Dal Bó & Fréchette, 2011; Fudenberg, Rand & Dreber, 2012) then allows the econometrician to estimate the importance of each strategy in the data, even though strategies are not observed.

The second type of FMM has proved uniquely useful in experimetrics because many economic experiments are expressly designed to distinguish among competing behavioural theories. In this setting, a mixture model does not merely signal that subjects are heterogeneous; it explicitly acknowledges that their heterogeneity may arise because different subjects use entirely different models of choice. By aligning the econometric framework with the experimental design, researchers can map observed choices directly to well-defined theoretical constructs.

Typically, FMMs are estimated using maximum likelihood, and a frequently encountered problem is non-uniqueness of the MLE (McLachlan & Peel, 2000). An alternative is Bayesian estimation, in which Bayesian updating rules are used to estimate the probability of a given subject using each decision rule. This approach claims to avoid the problem of non-uniqueness of the MLE. The Bayesian approach has been used by Houser, Keane and McCabe (2004) to classify subjects solving difficult decision problems into three types: near-rational; fatalist; and confused.

FMMs have stood the test of time because they address one of the central stylised facts of experimental data, that is, in the words of John Hey: "people are different".[2] Sometimes they differ only in the values of their parameters. Sometimes they differ in the behavioural rules they use. Sometimes both forms of heterogeneity are present.

[2] Conte et al. (2011), p.79.

The strength of finite mixture modelling is that it offers a tractable way of representing these differences while retaining a clear link to economic theory.

At the same time, FMMs require discipline. The number of types should not be chosen mechanically, and a better statistical fit is not enough if the resulting types have no economic interpretation. In experimetrics, the most successful applications of finite mixture modelling have been those in which the latent classes are either clearly interpretable as forms of unobserved heterogeneity, or are tied to well-defined behavioural rules. This is why the method has remained useful: it combined flexibility with interpretability, provided the researcher resists the temptation to let the mixture absorb every unexplained feature of the data.

14.7 Structural Estimation of Preference Parameters

By 'structural' models, we mean theoretical models that contain 'home-grown' preference parameters, such as risk aversion, probability weighting, inequity aversion, time discounting, and so on. These parameters typically display between-subject heterogeneity and investigators are typically interested in their distribution over the population.

A remarkably durable tool in experimental economics is the Multiple Price List (MPL). Its major appeal is that it is easy to understand. An MPL is simple to explain to subjects, simple to implement in the laboratory or in the field, and (sometimes but not always) simple to analyse. Early applications include the measurement of risk attitude (Binswanger, 1980), the elicitation of willingness to pay (Kahneman, Knetsch & Thaler, 1990), and the elicitation of individual discount rates (Coller & Williams, 1999). The most influential application in the risk-preference literature is the Holt and Laury design (Holt & Laury, 2002), which has become a standard workhorse in experimental economics.

The success of the MPL is, however, a useful example of a more general theme in experimetrics: what is convenient at the design stage is not necessarily innocuous at the estimation stage. In a standard MPL, subjects face an ordered sequence of binary choices. Under deterministic revealed preference, a subject with monotonic preferences should switch at most once from the safer to the riskier option. The location of the switch-point can then be mapped into an interval for the relevant preference parameter, for example a coefficient of risk aversion under an assumed parametric utility function. This is precisely what has made MPLs so popular: they convert a potentially complex preference-estimation problem into an apparently simple interval-data problem.

Yet the apparent simplicity is partly deceptive. Standard MPL data frequently contain multiple switching. Such behaviour may reflect errors, imprecision, indifference, confusion, or the fact that the list itself induces subjects to think about the rows jointly rather than as a sequence of independent decisions. The critique of list-based elicitation has been sharpened by Brown and Healy (2018), who show experimentally that the random problem selection mechanism may fail to be incentive compatible

when all decisions are displayed together as a single list, while incentive compatibility is restored when the same decisions are presented separately and in random order. Their result is especially important because it turns a common criticism of separated designs on its head: separated presentation produces more apparent inconsistency, but this may be because it reveals rather than creates the stochastic nature of preferences.

One response to the multiple-switching problem is the Switching Multiple Price List (sMPL), in which subjects are asked to report a single switch-point directly (Andersen, Harrison, Lau & Rutström, 2006). This design has the attractive feature of imposing monotonicity by construction. It is therefore particularly useful when the elicited preference parameter is not the main object of the study, but a control variable to be used in a subsequent behavioural analysis. In that setting, the sMPL offers a good compromise between subject comprehension, implementation costs, and econometric tractability.

The econometric interpretation of sMPL data is straightforward in the single-parameter case. Once a utility function has been specified, the reported switch-point implies an interval in which the subject's parameter must lie. Many researchers have proceeded by applying linear regression techniques to data on the midpoints of each interval, choosing arbitrary midpoints for open intervals. We emphasise that using midpoints does not lead to consistent estimation, and that the appropriate econometric tool is interval regression, which consistently estimates the distribution of the preference parameter in the population. This approach treats preference parameters as heterogeneous across subjects, while retaining a deterministic revealed preference interpretation within subjects. In other words, conditional on the subject's parameter, the observed switch-point is assumed to be chosen without error.

This last point is crucial. The sMPL *does not* solve the problem of stochastic choice; it 'kills' the information that would allow the econometrician to model it. With one switch-point per subject and one parameter of interest, there is no within-subject residual variation left to explain. The method is therefore well suited to estimating between-subject heterogeneity, but not to estimating within-subject variation. Put differently, the sMPL is useful precisely because it suppresses the type of inconsistency that stochastic models of choice are designed to explain.[3]

The restriction becomes more interesting, and more demanding econometrically, when more than one sMPL is used. Designs such as Tanaka, Camerer and Nguyen (2010) use more than one list in order to recover more than one preference parameter, for example risk aversion and probability weighting. In principle, this allows the researcher to move beyond Expected Utility and estimate the joint distribution of parameters in a non-EU model or a multi-parameter utility function. However, the data are no longer ordinary univariate interval data. A pair of switch-points identifies a region in the space of preference parameters, and this region will generally be non-rectangular. The likelihood contribution is therefore the probability mass of the assumed joint distribution over that region.

[3] This is the point at which the deterministic revealed preference assumption becomes most visible. One might say, adapting the well-known Monty Python line, that the assumption has sometimes been treated as if it were merely resting, when in many applications it is closer to a dead parrot.

Conte, Moffatt and Riddel (2026) develop an estimator for this setting, the Multivariate Heterogeneous Preference (MHP) estimator. In the bivariate case, with two sMPLs and two preference parameters, the estimator can be interpreted as a multivariate generalisation of interval regression. The generalisation is not mechanical, however. Because the regions implied by pairs of switch-points are irregular, the likelihood cannot be evaluated using standard interval-regression routines. Monte Carlo integration is required, and importance sampling is needed to deal with 'rare event sampling', that is with small or remote regions of the parameter space. This illustrates a broader lesson: the simplicity of the sMPL design may shift, rather than remove, the econometric burden.

The practical recommendation is therefore to distinguish sharply between two research objectives. If the purpose of eliciting risk attitude, discount rates, or other preference parameters is to obtain a parsimonious control for use in another model, then one or more sMPLs may be appropriate. In that case, the econometric analysis should respect the interval nature of the data, and, when more than one parameter is involved, should model the joint distribution of the parameters rather than replacing intervals by arbitrary midpoints.

If, instead, the object of the research is behaviour under risk itself (for example, testing Expected Utility, comparing alternative non-EU theories, estimating decision errors, or studying how behaviour changes with experience) then MPL or sMPL designs are too restrictive. A different design is required: a sequence of independent choice problems, preferably presented separately and not as an ordered list. Designs of this kind were used by Hey and Orme (1994) and subsequent studies, often with the aim of covering the Marschak-Machina triangle. Such data contain repeated choices at the subject level and are therefore suitable for fully structural stochastic choice models.

14.8 Competing Stochastic Specifications

In structural models of choice, the deterministic component of the model and the stochastic component of the model should be kept conceptually distinct. The deterministic component specifies how subjects evaluate the available options: for example, whether utility is CRRA or CARA, whether choices are evaluated according to Expected Utility, Rank-Dependent Expected Utility, or some other preference functional. The stochastic component specifies why observed choices may depart from the deterministic prediction. This second component is not a minor technical detail. Different stochastic specifications embody different views of what decision errors, inconsistencies, or repeated-choice variation mean.

Broadly, three different stochastic specifications have been especially important in the structural analysis of risky choice data: tremble models, Random Utility models often implemented through Fechnerian errors, and Random Preference models. All three introduce within-subject variation, but they do so in different ways. A tremble model attributes some choices to lapses, implementation mistakes, or random

responses. A Fechner model attributes stochastic choice to errors in the evaluation or comparison of utilities. A random preference model attributes stochastic choice to variation in the preference parameters themselves.

The tremble is the simplest of these stochastic specifications. The idea has a distant antecedent in Selten (1975)'s discussion of trembling-hand perfection in games: even if a player intends to choose optimally, there is a small probability that the intended action is not implemented. In experimental choice models, the same idea appears as a constant-probability error, or tremble. With some small probability, the subject is assumed not to follow the deterministic decision rule and instead to choose randomly. This type of error was used in the comparison of alternative theories of risky choice by Harless and Camerer (1994) and was later analysed explicitly by Moffatt and Peters (2001).

The appeal of the tremble is that it captures a form of mistake that is difficult to represent with a smooth error model. Subjects may press the wrong button, fail to pay attention, misunderstand a particular task, or make an occasional implementation error. Such mistakes may occurr even when the deterministic utility difference bewteen the two option is large. In this sense, the tremble can protect the estimation from treating a small number of anomalous choices as evidence aganinst the underlying preference model.

At the same time, a pure tremble model is too crude to be a complete theory of stochastic choice. It treats errors equally likely regardless of how close the alternatives are in utility terms. Therefore, it cannot capture the intuitive idea that subjects should be more likely to make mistakes when two options are nearly equivalent than when one option is clearly better. For this reason, the tremble has not mostly survived as a stand-alone model, but as an auxiliary component that can be added to Fechner, Random Preference, or other structural spcifications. Its role is to capture occasional random responses, while the main stochastic specification captures systematic within-subject variation.

The Fechner model first appeared long ago in the Psychometrics literature (Fechner, 1860), and first appeared in the experimetrics literature with Hey and Orme (1994). This model assumes that the subject has stable preference parameters but makes errors in evaluating or comparing the utilities of the available options. In the binary-choice case, the probability of choosing one lottery rather than the other is usually modelled as an increasing function of the difference between their deterministic utilities. The closer the two options are to in utility terms, the more likely the subject is to choose the option with lower deterministic utility.

The Random Preference model (Loomes & Sugden, 1998; Loomes, Moffatt & Sugden, 2002) assumes instead that the preference parameters themselves vary from one decision to the next. The subject may therefore behave as if they were more or less risk averse from one task to the next. This approach has an appealing behavioural interpretation, since inconsistency is treated as preference variation rather than as computational error.

This brings us to a central question in the stochastic specification of structural models: how should the econometrician interpret within-subject variation? Is it best understood as random implementation error, as noise in utility evaluation, or

as variation in preferences themselves? The debate between Fechner and Random Preference specifications has been central to experimetrics for almost three decades, and it is to this debate that we now turn.

The distinction between these two approaches has been sharpened by the monotonicity critique of Random Utility models. The basic point is simple. If a higher value of the risk-aversion parameter makes the risky lottery less attractive, then the probability of choosing that lottery should decrease as risk aversion increases. Standard Fechner specifications based on CRRA or CARA expected utility differences need not satisfy this property. The utility difference between two lotteries may be non-monotone in the preference parameter, and therefore the implied choice probability may also be non-monotone. This can create identification problems and may distort maximum-likelihood estimates of risk aversion (Wilcox, 2011; Apesteguia & Ballester, 2018).

This critique is important, but it should not be interpreted as a definitive rejection of the Fechner approach and there are studies which are trying to limit the relevance of the non-monotonicity critique (Conte & Hey, 2026). Anyhow, the Fechner model remains one of the most intuitive stochastic representations of choice: subjects attempt to choose the alternative with the higher deterministic utility, but make errors in the evaluation or implementation of that comparison.

For this reason, the Fechner model has survived the test of time. Its theoretical limitations are now better understood, but its combination of transparency, tractability and flexibility has made it the workhorse of structural estimation in experimetrics. Random Preference models provide an important benchmark and avoid some of the monotonicity problems associated with standard Random Utility specifications. However, they are typically harder to estimate, especially when the lotteries in the choice problems are complex, or when the model of interest departs from Expected Utility and includes additional preference parameters, such as probability weighting, loss aversion or ambiguity attitude. In practice, the two approaches should therefore be seen as complementary rather than as mutually exclusive. Fechner models offer a tractable way of modelling computational error; Random Preference models offer a disciplined way of modelling preference variation. Which one is appropriate depends on the research question, the experimental design, and the complexity of the structural model.

There is also a modelling asymmetry in some comparisons between Random Utility and Random Preference specifications. The benchmark Fechner model is often implemented as a representative-agent model, with a common preference parameter and an additive stochastic error. In this specification, individuals may make different choices because they receive different error draws, but they are assumed to share the same structural preference parameters. Random Preference models, by contrast, place the stochastic component directly on the preference parameter, and are therefore closer to a model of preference heterogeneity or preference instability. The comparison is therefore not only between two error specifications, but also between two different ways of representing heterogeneity.

This observation points to an important development in the structural estimation of experimental data. Early applications often relied on representative-agent models,

or on pooled specifications in which all subjects were assumed to share the same preference parameters. A first step beyond this approach is to allow the structural parameters to vary with observed individual characteristics. For example, risk aversion, probability weighting, loss aversion, or discounting parameters may be specified as functions of demographic variables, treatment dummies, or other observed covariates. This captures observed heterogeneity, but still assumes that individuals with the same covariates have the same structural parameters.

A further step is to allow for both observed and unobserved heterogeneity. In random-coefficient structural models, the preference parameters are treated as draws from a population distribution whose moments may depend on observable characteristics. In the context of choices under risk, Botti, Conte, Di Cagno and D'Ippoliti (2008) estimate random-coefficient structural probit models in which risk attitude and non-EU parameters vary with both demographic variables and unobserved individual components. This makes it possible to estimate not only average preference parameters, but also the dispersion of preferences in the population and the correlation between different preference dimensions.

This richer representation comes at an econometric cost. Once structural parameters are random, the individual likelihood contribution must be integrated over their population distribution. When more than one parameter is allowed to vary, as in models with both risk aversion and probability weighting, the likelihood typically contains multidimensional integrals with no closed-form solution. This is precisely the setting in which maximum simulated likelihood becomes useful: simulation replaces analytically intractable integrals by averages over draws from the assumed distribution of random coefficients.

As discussed above, finite mixture models provide a natural way of representing discrete heterogeneity. In structural models of choice under risk, this idea can be used to allow not only for heterogeneity in parameter values, but also for heterogeneity in the preference functional itself. Instead of assuming that all subjects are described by the same preference functional, the econometrician allows for distinct latent types. Some subjects may be Expected Utility maximisers, while others may be Rank-Dependent Expected Utility maximisers; alternatively, different types may correspond to different behavioural rules. Conte et al. (2011) develop this approach for choice under risk by estimating a mixture of EU and RDEU types, while also allowing for heterogeneity in the parameters within each type and for within-subject stochastic choice through Fechner errors and trembles.

The general lesson is that the stochastic specification of a structural model and the treatment of heterogeneity cannot be separated. Fechner errors, random preferences, random coefficients, and finite mixtures are not merely alternative technical devices. They embody different views of what the unexplained variation in experimental data represents: computational error, instability of preferences, continuous population heterogeneity, or discrete heterogeneity in decision types. As the representation of heterogeneity becomes richer, the econometric problem becomes more demanding. This is why simulation-based estimation methods, and in particular maximum simulated likelihood, have become central tools in experimetrics.

14.8.1 Maximum Simulated Likelihood

Maximum Simulated Likelihood (MSL) is not itself a behavioural model. It is an estimation method that became important in experimetrics when behavioural models became richer. Once subjects are allowed to differ in unobserved ways, and once more than one structural parameter is allowed to vary across subjects, the likelihood often involves integrals that cannot be evaluated analytically. In similar cases, MSL provides a practical way of estimating such models by replacing these intractable integrals with simulated averages (Train, 2009; Moffatt, 2015).

The method is particularly useful in models with random coefficients. In the structural estimation of choice under risk, for example, the researcher may want to allow risk aversion, probability weighting, loss aversion, or error propensity to vary across individuals. If only one parameter is random, numerical integration may still be feasible. If two or more parameters are random, especially when they are correlated, the computational burden increases quickly. MSL became attractive because it allowed researchers to estimate models with several dimensions of unobserved heterogeneity without abandoning the likelihood framework.

A key example is the estimation of models in which both risk aversion and probability weighting are allowed to vary. Conte et al. (2011), for example, use MSL in a setting in which subjects may differ not only in the parameters of their preference functional, but also in the preference functional itself. Similarly, Von Gaudecker, Van Soest and Wengström (2011) assume four dimensions of between-subject heterogeneity: risk aversion, loss aversion, preference toward the timing of uncertainty resolution, propensity to choose randomly. The same general logic has been used to allow for heterogeneity in discount rates (Andersen, Harrison, Lau & Rutström, 2008), complexity aversion (Moffatt, Sitzia & Zizzo, 2015), inequity aversion (Gauriot, Heger & Slonim, 2020), and propensity to update beliefs (Henckel, Menzies, Moffatt & Zizzo, 2022).

The attraction of MSL is therefore both conceptual and practical. Conceptually, it enables the econometrician to move beyond the representative-agent model without estimating a separate model for each subject. The population is described by a distribution of preference parameters, and each subject's choices contribute to the estimation of that distribution. Practically, simulation makes it possible to approximate likelihood contributions that would otherwise require complicated multidimensional integration.

The use of simulation also made the choice of draws important. Simple pseudo-random draws can work, but they are required in large numbers to achieve stable estimates. Quasi-random sequences, such as Halton draws, have become popular because they cover the relevant space more evenly than independent random draws. In practice, this can reduce simulation noise and improve the precision of the simulated likelihood for a given number of draws. This is one reason why Halton draws became a standard tool in simulated maximum likelihood estimation, especially after their systematic use in the discrete-choice literature (Train, 2009).

The joint estimation of more than one preference parameter has also proved to be substantively important. In time-preference experiments, for example, allowing for

risk aversion rather than imposing risk neutrality can substantially reduce estimated discount rates (Andersen et al., 2008). In dictator-game data, allowing for risk aversion can similarly reduce estimates of inequity aversion (Gauriot et al., 2020). These examples show that MSL is not merely a computational convenience. By making joint structural estimation feasible, it can change the economic interpretation of experimental behaviour.

At the same time, MSL is not costless. The simulated likelihood is only an approximation to the true likelihood, and the quality of the approximation depends on the number and type of draws. The objective function may be noisy, especially when too few draws are used. Models with many random parameters may also be sensitive to starting values, distributional assumptions, and restrictions on the correlation structure among parameters. These practical difficulties explain why MSL has been most successful in applications where the additional heterogeneity has a clear behavioural interpretation and where the number of random parameters is kept manageable.

Thus, MSL has stood the test of time because it solved a problem that became central in experimetrics: how to estimate structural models that allow subjects to differ in unobserved ways. It provided a bridge between simple pooled models and fully individual-level estimation. It also paved the way for later Bayesian hierarchical approaches, which address similar questions about heterogeneity but within a different inferential framework.

14.8.2 Bayesian Hierarchical Models

An alternative approach to MSL that is currently gaining traction is the Bayesian hierarchical model. Examples of the use of these in Experimetrics can be traced back to Nilsson, Rieskamp and Wagenmakers (2011), although we note their relative obscurity until the beginning of the 2020s.[4] Here, instead of integrating over the uncertainty in each participant's individual-level parameters, these parameters are jointly estimated using Bayesian techniques alongside the parameters governing the distribution of the population. For example Gao, Harrison and Tchernis (2023) showcases some advantages of Bayesian hierarchical models in estimating risk preferences from binary choice data in economic experiments. Since participants' parameters are *estimated* rather than integrated out (as in MSL), economically meaningful quantities can be computed for specific participants, such as certainty equivalents measuring a participant's welfare gain associated with a lottery. Furthermore, since the full posterior distribution of the model's parameters is computed through simulation,

[4] This is perhaps a result of steady improvements in computational power, alongside improvements in software implementations needed to simulate posterior distributions reliably and quickly.

expressions of uncertainty of these quantities are easily computed simply by applying the desired function (e.g. the certainty equivalent) to these posterior draws.[5]

14.9 Econometric Modelling of Behaviour in Games

Most of the techniques covered in this chapter are techniques that Experimental Economists have been 'borrowed' from the mainstream Econometrics literature. The Experimetrics of behaviour in games is a good topic on which to end because the econometric techniques used for these applications tend, in contrast, to be ones that have been developed by Experimental Economists in direct response to their own needs.

14.9.1 Quantal Response Equilibrium

The Quantal Response Equilibrium (QRE) model was developed by McKelvey and Palfrey (1995). It is based on the idea that best responses are not played with certainty. Each player is assumed to calculate the expected value of each of her available actions given her (correct, in equilibrium) belief about her opponent's choice probabilities, and she attempts to respond optimally, but makes random errors in the process. Thus QRE replaces the perfectly rational expectations equilibrium embodied in Nash equilibrium with an imperfect, or 'noisy', rational expectations equilibrium. The principle of an equilibrium is maintained by assuming that players estimate expected payoffs in an unbiased way.

Estimating QRE models presents somewhat different challenges to other typical likelihood-based estimations implemented in Experimetrics. Whereas other applications, if computationally intensive, will be so due to difficulty in computing and maximizing the likelihood (or simulating the associated posterior). In the other hand, since QRE includes a fixed point condition, it is the computation of this condition that is burdensome in most applications. Bland and Turocy (2025) discusses this problem in detail and methods to overcome it in typical applications.

14.9.2 Learning Models

It is typical for subjects to engage in a sequence of tasks, and the investigator is often interested in the ways in which behaviour changes between tasks. The most obvious way of allowing for such learning processes is by including task number

[5] Using the MSL approach, computing such expressions of uncertainty would require either bootstrapping (which is computationally costly) or using the delta method (which is only asymptotically valid).

as an explanatory variable. This is routinely done in the analysis of public goods experiments (e.g. Bardsley & Moffatt, 2007), in which the usual pattern is subjects moving towards Nash behaviour, that is, learning to be selfish. Another type of learning process captured by task number is when a tremble probability is assumed to decay towards zero as the experiment progresses. An example is Loomes et al. (2002) who also discovered a strong tendency for subjects to move towards Expected Utility maximisation in the course of a risky choice experiment.

In the contexts of these examples, learning is only about the task, not about the behaviour of other players, nor even about the outcomes of previous tasks. That is why, in those situations, the modelling of a learning process simply involved allowing particular parameters to depend in some way on the task number.

The situation in repeated interactive games is very different, because subjects typically learn directly about the behaviour of others, and also about the type of strategies that are most profitable for themselves. The process of learning about other players is complicated by the fact that other players' behaviour changes as they, too, gain experience. A comprehensive model of learning should therefore incorporate the effects of the player's past pay-offs, and also the effects of past choices by other players.

A number of learning models have appeared in the literature with the aim of meeting some or all of these objectives, including Directional learning (Selten & Stoecker, 1986), Reinforcement Learning (Erev & Roth, 1998), Belief Learning (Cheung & Friedman, 1998) and Experience Weighted Attraction (EWA) (Camerer & Hua Ho, 1999).

Like risk models, learning models are models of individual behaviour, and so care needs to be taken in their treatment of between-subject heterogeneity if data are pooled for estimation. Wilcox (2006) shows that ignoring this heterogeneity in EWA models leads to a severe estimation bias that favours Reinforcement Learning over Belief Learning. Appropriately modelling the heterogeneity, whether through MSL (as in Cason, Friedman & Hopkins, 2010) or a Bayesian hierarchical model (as in Bland, 2025b, Chapter 10), is therefore crucial.

14.10 Summary and Conclusion

An interesting feature of the research area covered in this chapter is that many aspects of experimental design, and econometric procedures, that are favoured by Experimental Economists, and have consequently stood the test of time, are not the most suitable techniques from a theoretical perspective.

An obvious example is that researchers who are concerned about order effects have a preference for between-sample tests, despite the clear advantages of within-sample tests in terms of both efficiency and cost. For another example, the common preference for non-parametric tests resulting from concerns about non-normality is surely misguided since a bootstrapped parametric test would be more powerful while also being valid. Yet another straighforward example where the most efficient

approach appears not to be the preferred appraoch arises in the analysis of risk attitude. From a theoretical perspective, the most efficient way of eliciting risk attitude is to ask for a certainty equivalent, but for various behavioural reasons (e.g., reporting a certainty equivalent is a task that is unfamiliar to most subjects) this method of elicitation is thought to distort true risk attitudes, and choice between lotteries is the preferred elicitation method. As a final example, in the context of structural modeling, while the random preference model is possibly the most elegant and intuitive econometric specification, the Fechner model is preferred for practical reasons. In all of these examples, behavioural considerations appear to be pulling in the opposite direction to theoretical considerations. Behavioural considerations usually come out on top.

We have identified two broad types of Economic Experiment. The first is theory-testing experiments, whose aim is to test the internal logic of economic models, studying comparative static effects, studying the validity of theoretical isomorphisms, comparing the efficiency of institutions, studying adaptive adjustment processes towards equilibrium. For such experiments, the designs are relatively complex since they are typically based on induced valuation methodology (Smith, 1962), which is essentially a means of removing the effects of subject preferences so that the focus is on the theory under test.

The second type of experiment might be labelled 'measurement' experiments. The broad goal of this literature is to use usually relatively simple experimental designs (such as lottery choice or allocation decisions) to measure characteristics subjects bring into the laboratory with them, especially preference parameters and cognitive parameters.

As soon as we move from the first type of experiment to the second, we need to start dealing with between-subject heterogeneity, and this calls for more advanced econometric techniques. For this reason, we, as econometricians, are more interested in the second type of experiment than the first, and hence these types of experiment have been the main focus of this chapter.

There are broadly two types of heterogeneity seen in experimental data: discrete and continuous. For discrete heterogeneity, the Finite Mixture Modelling approach has been extremely useful in the Experimetrics literature ever since El-Gamal and Grether (1995). This approach allows subjects to differ completely in their motivations, which is apparently what is required when analysing experimental data, and usually provides answers to the research questions of interest.

To deal with continuous heterogeneity, one relatively straightforward approach that we have advocated is the switching multiple price list (sMPL). This technique has been used extensively in the literature, to account for heterogeneity in many different preference parameters. We have stressed the point that the sMPL is adequate if the objective is to obtain estimates of subjects' preference parameters as an explanatory factor in some other aspect of behaviour. But if the focus of the investigation is on the preference parameter itself, a more elaborate design consisting of a sequence of independent choice problems is recommended, and consequently a more sophisticated set of econometric techniques are required.

The two econometric approaches we have presented are Maximum Simulated Likelihood and Bayesian hierarchical models. Both of these approaches are proving to be very promising. The first is particularly useful for estimation of models with more than one preference parameter (see e.g., Von Gaudecker et al., 2011). The second is particularly useful when economically meaningful quantities are required for each subject. Which of the two approaches 'stands the test of time' is a question for the future.

References

Alekseev, A. (2025). The (statistical) power of incentives. *Journal of the Economic Science Association*, 1–18.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: journal of the Econometric Society*, 503–546.

Andersen, S., Harrison, G. W., Lau, M. I. & Rutström, E. E. (2006). Elicitation using multiple price list formats. *Experimental Economics*, *9*(4), 383–405.

Andersen, S., Harrison, G. W., Lau, M. I. & Rutström, E. E. (2008). Eliciting risk and time preferences. *Econometrica*, *76*(3), 583–618.

Andreoni, J. (1988, December). Why free ride? : Strategies and learning in public goods experiments. *Journal of Public Economics*, *37*(3), 291-304. Retrieved from https://ideas.repec.org/a/eee/pubeco/v37y1988i3p291-304.html doi: None

Apesteguia, J. & Ballester, M. A. (2018). Monotone stochastic choice models: The case of risk and time preferences. *Journal of Political Economy*, *126*(1), 74–106.

Atkinson, A. C. (1996). The usefulness of optimum experimental designs. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 59–76.

Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C. & Sugden, R. (2009). *Experimental economics: Rethinking the rules*. Princeton University Press.

Bardsley, N. & Moffatt, P. G. (2007). The experimetrics of public goods: Inferring motivations from contributions. *Theory and Decision*, *62*(2), 161–193.

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis. *Papers Imp. Acad. Sci.*, *5*, 175–192.

Binswanger, H. P. (1980). Attitudes toward risk: Experimental measurement in rural india. *American Journal of Agricultural Economics*, *62*(3), 395–407.

Bland, J. R. (2025a). Optimizing experiment design for estimating parametric models in economic experiments. *Available at SSRN 4566855*.

Bland, J. R. (2025b). Structural bayesian techniques for experimental and behavioral economics. *Available at SSRN 5241268*.

Bland, J. R. & Turocy, T. L. (2025). Quantal response equilibrium as a structural model for estimation: The missing manual. *Games and Economic Behavior*.

Bosch-Domènech, A., Montalvo, J. G., Nagel, R. & Satorra, A. (2010). A finite mixture analysis of beauty-contest data using generalized beta distributions. *Experimental Economics*, *13*(4), 461–475.

Bosman, R. & Van Winden, F. (2002). Emotional hazard in a power-to-take experiment. *The Economic Journal*, *112*(476), 147–169.

Botti, F., Conte, A., Di Cagno, D. T. & D'Ippoliti, C. (2008, March). Risk attitude in real decision problems. *The B.E. Journal of Economic Analysis & Policy*, *8*(1), 1–32.

Brown, A. L. & Healy, P. J. (2018). Separated decisions. *European Economic Review*, *101*, 20–34. doi: 10.1016/j.euroecorev.2017.09.014

Bruhin, A., Fehr, E. & Schunk, D. (2019). The many faces of human sociality: Uncovering the distribution and stability of social preferences. *Journal of the European Economic Association*, *17*(4), 1025–1069.

Camerer, C., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., ... others (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436.

Camerer, C. & Hua Ho, T. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, *67*(4), 827–874.

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø. & Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, *97*(3), 818–827.

Cason, T. N., Friedman, D. & Hopkins, E. (2010). Testing the tasp: An experimental investigation of learning in games with unstable equilibria. *Journal of Economic Theory*, *145*(6), 2309–2331.

Charness, G., Gneezy, U. & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8.

Chaudhuri, P. & Mykland, P. A. (1995). On efficient designing of nonlinear experiments. *Statistica Sinica*, 421–440.

Cheung, Y.-W. & Friedman, D. (1998). A comparison of learning and replicator dynamics using experimental data. *Journal of Economic Behavior & Organization*, *35*(3), 263–280.

Clark, J. (2002). House money effects in public good experiments. *Experimental Economics*, *5*(3), 223–231.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. routledge.

Coller, M. & Williams, M. B. (1999). Eliciting individual discount rates. *Experimental Economics*, *2*(2), 107–127.

Conlisk, J. (1989). Three variants on the allais example. *The American Economic Review*, 392–407.

Conte, A. & Hey, J. D. (2026, May). *REDRUM: Robust Estimation and Design for the Random Utility Model* (MPRA Paper No. 129162). University Library of Munich, Germany. Retrieved from https://ideas.repec.org/p/pra/mprapa/129162.html doi: None

Conte, A., Hey, J. D. & Moffatt, P. G. (2011). Mixture models of choice under risk. *Journal of Econometrics*, *162*(1), 79–88.

Conte, A., Levati, M. V. & Montinari, N. (2019, February). Experience in public goods experiments. *Theory and Decision*, *86*(1), 65-93. doi: 10.1007/s11238 -018-9670-z

Conte, A. & Moffatt, P. G. (2014). The econometric modelling of social preferences. *Theory and Decision*, *76*(1), 119–145.

Conte, A., Moffatt, P. G. & Riddel, M. (2026). The Multivariate Heterogeneous Preference estimator for Switching Multiple Price List data. *Experimental Economics*, 1–25. doi: 10.1017/eec.2026.10046

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: journal of the Econometric Society*, 829–844.

Dal Bó, P. & Fréchette, G. R. (2011). The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review*, *101*(1), 411–429.

Dong, D., Chung, C. & Kaiser, H. M. (2004). Modelling milk purchasing behaviour with a panel data double-hurdle model. *Applied Economics*, *36*(8), 769–779.

Duan, N., Manning, W. G., Morris, C. N. & Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of business & economic statistics*, *1*(2), 115–126.

Eckel, C. C. & Grossman, P. J. (2000). Volunteers and pseudo-volunteers: The effect of recruitment method in dictator experiments. *Experimental Economics*, *3*(2), 107–120.

El-Gamal, M. A. & Grether, D. M. (1995). Are people bayesian? uncovering behavioral strategies. *Journal of the American Statistical Association*, *90*(432), 1137–1145.

Ellingsen, T. & Johannesson, M. (2004). Promises, threats and fairness. *The Economic Journal*, *114*(495), 397–420.

Engel, C. & Moffatt, P. G. (2012). Estimation of the house money effect using hurdle models. *MPI Collective Goods Preprint*(2012/13).

Engelmann, D. & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American economic review*, *94*(4), 857–869.

Erev, I. & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, 848–881.

Fechner, G. T. (1860). *Elemente der psychophysik*. Leipzig, Germany: Breitkopf und Härtel.

Fisher, R. (1926). The arrangement of field experiments. *J. Minist. Agric.*, *33*, 503–513.

Forsythe, R., Horowitz, J. L., Savin, N. E. & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, *6*(3), 347–369.

Fréchette, G. R. (2012). Session-effects in the laboratory. *Experimental Economics*, *15*(3), 485–498.

Fudenberg, D., Rand, D. G. & Dreber, A. (2012). Slow to anger and fast to forgive: Cooperation in an uncertain world. *American Economic Review*, *102*(2),

720–749.

Gächter, S. & Renner, E. (2010). The effects of (incentivized) belief elicitation in public goods experiments. *Experimental Economics*, *13*(3), 364–377.

Gao, X. S., Harrison, G. W. & Tchernis, R. (2023). Behavioral welfare economics and risk preferences: A bayesian approach. *Experimental Economics*, *26*(2), 273–303.

Gauriot, R., Heger, S. A. & Slonim, R. (2020). Altruism or diminishing marginal utility? *Journal of Economic Behavior & Organization*, *180*, 24–48.

Grether, D. M. & Plott, C. R. (1979). Economic theory of choice and the preference reversal phenomenon. *The american economic review*, *69*(4), 623–638.

Ham, J. C., Kagel, J. H. & Lehrer, S. F. (2005). Randomization, endogeneity and laboratory experiments: The role of cash balances in private value auctions. *Journal of Econometrics*, *125*(1-2), 175–205.

Harless, D. W. & Camerer, C. F. (1994). The predictive utility of generalized expected utility theories. *Econometrica*, *62*(6), 1251–1289. doi: 10.2307/2951749

Harrison, G. W. & Rutström, E. E. (2009). Expected utility theory and prospect theory: One wedding and a decent funeral. *Experimental economics*, *12*(2), 133–158.

Henckel, T., Menzies, G. D., Moffatt, P. G. & Zizzo, D. J. (2022). Belief adjustment: A double hurdle model and experimental evidence. *Experimental Economics*, *25*(1), 26–67.

Hey, J. D. & Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica: Journal of the Econometric Society*, 1291–1326.

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

Holt, C. A. & Laury, S. K. (2005). Risk aversion and incentive effects: New data without order effects. *American Economic Review*, *95*(3), 902–904.

Holt, C. A. & Sullivan, S. P. (2023). Permutation tests for experimental data. *Experimental Economics*, *26*(4), 775–812.

Houser, D., Keane, M. & McCabe, K. (2004). Behavior in a dynamic decision problem: An analysis of experimental evidence using a bayesian type classification algorithm. *Econometrica*, *72*(3), 781–822.

Huber, J. & Zwerina, K. (1996). The importance of utility balance in efficient choice designs. *Journal of Marketing research*, *33*(3), 307–317.

Isoni, A., Loomes, G. & Sugden, R. (2011). The willingness to pay—willingness to accept gap, the "endowment effect," subject misconceptions, and experimental procedures for eliciting valuations: Comment. *American Economic Review*, *101*(2), 991–1011.

Johnson, C., Baillon, A., Bleichrodt, H., Li, Z., Van Dolder, D. & Wakker, P. P. (2021). Prince: An improved method for measuring incentivized preferences. *Journal of Risk and Uncertainty*, *62*(1), 1–28.

Kagel, J. H. & Levin, D. (1986). The winner's curse and public information in common value auctions. *The American Economic Review*, 894–920.

Kahneman, D., Knetsch, J. L. & Thaler, R. H. (1990). Experimental tests of the endowment effect and the coase theorem. *Journal of Political Economy*, *98*(6), 1325–1348.

Lipsey, R. G. (1979). *An introduction to positive economics* (6th ed.). London: Weidenfeld and Nicolson.

Loomes, G., Moffatt, P. G. & Sugden, R. (2002). A microeconometric test of alternative stochastic theories of risky choice. *Journal of Risk and Uncertainty*, *24*(2), 103–130.

Loomes, G. & Sugden, R. (1998). Testing different stochastic specificationsof risky choice. *Economica*, *65*(260), 581–598.

McKelvey, R. D. & Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and economic behavior*, *10*(1), 6–38.

McLachlan, G. J. & Peel, D. (2000). *Finite mixture models*. John Wiley & Sons.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157.

Moffatt, P. G. (2007). Optimal experimental design in models of decision and choice. *Measurement in Economics*, *357*.

Moffatt, P. G. (2015). *Experimetrics: Econometrics for experimental economics*. Macmillan International Higher Education.

Moffatt, P. G. (2019). Data analysis. In *Handbook of research methods and applications in experimental economics* (pp. 57–82). Edward Elgar Publishing.

Moffatt, P. G. (2021). Experimetrics: a survey. *Foundations and Trends in Econometrics*, *11*(1-2), 1–152.

Moffatt, P. G. (2025). Experimetrics. In A. N. S-H Chuah R. Hoffmann (Ed.), *Elgar encyclopedia of behavioural and experimental economics* (pp. 210–212). Edward Elgar.

Moffatt, P. G. (2026). Experimetrics. In B. Kebede (Ed.), *Encyclopedia of experimental social science*. Edward Elgar.

Moffatt, P. G. & Peters, S. A. (2001). Testing for the presence of a tremble in economic experiments. *Experimental Economics*, *4*(3), 221–228. doi: 10.1023/A:1013265203635

Moffatt, P. G., Sitzia, S. & Zizzo, D. J. (2015). Heterogeneity in preferences towards complexity. *Journal of Risk and Uncertainty*, *51*(2), 147–170.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of econometrics*, *32*(3), 385–397.

Nilsson, H., Rieskamp, J. & Wagenmakers, E.-J. (2011). Hierarchical bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*, *55*(1), 84–93.

Oshchepkov, A. & Shirokanova, A. (2022). Bridging the gap between multilevel modeling and economic methods. *Social Science Research*, *104*, 102689.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. I. On the dissection of asymmetrical frequency-curves. *Philosophical Transactions of the Royal Society of London. A*, *185*, 71–110. doi: 10.1098/rsta.1894.0003

Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, *4*(1), 25–55. doi:

10.1007/BF01766400

Selten, R. & Stoecker, R. (1986). End behavior in sequences of finite prisoner's dilemma supergames a learning theory approach. *Journal of Economic Behavior & Organization*, *7*(1), 47-70. Retrieved from https://www.sciencedirect.com/science/article/pii/0167268186900211 doi: https://doi.org/10.1016/0167-2681(86)90021-1

Siegel, S. & Castellan. (1988). *Nonparametric statistics for the social sciences*. New York: McGraw-Hill.

Silvey, S. D. (1980). Sequential designs. In *Optimal design: An introduction to the theory for parameter estimation* (pp. 62–68). Springer.

Smith, V. L. (1962). An experimental study of competitive market behavior. *Journal of Political Economy*, *70*(2), 111–137.

Tanaka, T., Camerer, C. F. & Nguyen, Q. (2010). Risk and time preferences: Linking experimental and household survey data from vietnam. *American Economic Review*, *100*(1), 557–571.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, 24–36.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge University Press.

Von Gaudecker, H.-M., Van Soest, A. & Wengström, E. (2011). Heterogeneity in risky choice behavior in a broad population. *American Economic Review*, *101*(2), 664–694.

Wilcox, N. T. (2006). Theories of learning in games and heterogeneity bias. *Econometrica*, *74*(5), 1271–1292.

Wilcox, N. T. (2011). Stochastically more risk averse: A contextual theory of stochastic discrete choice under risk. *Journal of Econometrics*, *162*(1), 89–104.

Zhang, L. & Ortmann, A. (2013). Exploring the meaning of significance in experimental economics. *UNSW Australian School of Business Research Paper*(2013-32).