Experimental Economics Across Subject Populations

Guillaume R. Fréchette
New York University
August 2014

Abstract

This chapter reviews and discusses the results of experiments using non-standard subjects. In particular, experiments using non-human animals, people living in token economies, children, the elderly, demographically varied samples, and professionals are considered. Investigating such diverse subject pools offers a perspective on the generalizability of findings with the standard subject pool and reveals what results apply broadly, which behaviors are learned, and if some behavior can be un-learned or mitigated via selection or experience. The chapter proceeds subject pool by subject pool. Each section also provides reasons why specific subject pools are interesting to study, as well as some observations of a methodological nature concerning that group.

A final section discusses behavior across subject pools in a given game or choice task, including: choices over goods, choices over lotteries, the estimation of risk and time preferences, bargaining games, public good games, trust games, and beauty contest games. Although the decision tasks and games are quite diverse, it appears that most qualitative comparative statics results are robust across subject populations. However, some differences emerge as well, but these are almost exclusively with respect to quantitative results.

* Fréchette: Department of Economics, New York University, New York NY 10012 (e-mail: frechette@nyu.edu). I wish to thank Jacopo Perego for his tremendous help and comments as well as Guillaume Hollard, John Kagel, Kenway Louie, Andrew Schotter, Emanuel Vespa, and Alistair Wilson for their feedback. I gratefully acknowledge the support of NSF via grants SES-0924780 and SES-1225779 as well as support from the Center for Experimental Social Science (CESS), and the C.V. Starr Center. Its contents are solely the responsibility of the author and do not necessarily represent the official views of the NSF.

I. Introduction

This review will cover results from experiments on different subject pools, none of which are the usual sample of undergraduate students: infrahumans (rats, pigeons, and monkeys), people living in a token economy (mostly mental institutions), children, the elderly, subjects that are representative of larger populations, and professionals.[1] What can we learn from studying rats or patients with psychiatric issues? What is the interest in reviewing results from experiments on such different subjects? By analyzing experimental data from different subject pools, we can assess to what extent behavior across species or groups features similar patterns. There are similarities in behavior across humans despite many differences in background, as well as between humans and other animals. Suppose, for instance, that one observed a puzzling behavior for the first time, say something such as hyperbolic discounting, in an experiment with the standard subject pool of undergraduate college students. One reaction could be to think that this is an artifact of the method of discovery: either the subject pool, the size of the incentives, etc. However, suppose a similar behavior was observed in rats, pigeons, children, patients suffering from depression, etc. Then, it would seem to build a case, from the weight of evidence, that this is not simply an artifact of the standard economic experimental methods and subject pool, but rather a robust phenomenon. This, I think, is one of the crucial reasons for which experiments with non-standard subject pools are interesting: they allow for a better test of the robustness of our theories and findings.[2]

The counterpart to the point above is that it also allows us to discover specific ways in which the behavior of these groups differs. This is also interesting in itself. Indeed, learning how specific groups differ from each other helps us gain a deeper understanding of the nature of the group itself. For example, studying children can shed light on questions such as nature vs. nurture. For behaviors that are learned, it offers an opportunity to understand when they appear. Understanding how preferences are shaped by ageing may be key to designing optimal policies.

[1] I use infrahuman in the narrow sense of non-human animal.

[2] A related but distinct issue is whether subjects who volunteer for experiments are different from others who do not. The issue of volunteer artifact in an experiment on electricity demand (how it is affected by changes in prices and other factors) is explored in Kagel et al. (1979). See also Falk et al. (2013) described in Section VI of this chapter.

Despite the ways in which these groups are interesting, two samples stand out as more obviously of interest to economists: the representative sample (i.e. representative of the population at large) and professionals (meaning people who work in an area of interest). Although most economic models typically do not specify who they apply to, the groups on which one would think it makes most sense to test, evaluate, or estimate economic models on, are a representative sample of the population or samples of specialists at a certain task. For example, the representative sample could be of particular interest for analyzing savings, labor supply decisions, voting, and the like, while a sample of specialists could be used to study, for example, the behavior of professional traders in financial markets. In that sense, these two samples are the most natural to study. Yet, these two samples do also present disadvantages, and, conversely, the other samples, which are somewhat more removed from our immediate interest, present advantages.

There are four main disadvantages to studying representative samples and professionals. These are costs, availability, replicability, and the limits to control. Costs are usually an issue because the opportunity cost of participating in an experiment for a student is certainly lower than that of an average member of a large population or, even more, of a professional. For this reason, providing appropriate incentives when using these non-standard samples can turn out to be particularly expensive, undermining the researcher's ability to gather a sufficient amount of data.

A second issue is availability. It is usually more difficult, to have access to a representative sample, or a pool of professionals, especially compared with how easy it is to recruit students. In fact, this explains why some of the research with professionals uses such unusual professionals: sportscard traders, fruit pickers, tree planters, soccer players, etc. These are not necessarily the typical professionals we have in mind and their use often has more to do with the connections or personal interests of the authors, which make it easier for them to have access to these people.

These first two factors, cost and availability, combine together to form a much more important problem: the replicability of experiments with these samples is severely reduced. With a standard experiment, if one doubts the robustness of a particular result and thinks the result is caused by some of the details of the implementation, be it the instructions, the interface, or the specific parameters used; it is relatively easy to conduct

an experiment that varies the aspects of concern. For topics that become popular, the tradition of repeating one's control can by chance reveal important details. An example of this is Charness, et al. (2004) where the authors added in the player's instructions a simple table to summarize the payoffs of the manager and employee in a gift-exchange game. They found that this simple change substantially reduced the amount of gift-exchange, suggesting that part of the previous results were simply due to misunderstanding the implications of the player's actions. The limited replicability of experiments with representative samples and professionals is a non-trivial shortcoming.

Finally, using representative samples or professionals often means having less control on the experimental environment. Subjects who are unaccustomed to written instructions and to abstractions can have a tendency to understand the environment they are presented with differently from the way it is intended, introducing noise in the observed data. In particular, there is evidence of professionals behaving as if they are in their professional environment even when this is not appropriate. Although this is instructive, as it speaks to the importance of those aspects of their professional environment, it can be, at times, a nuisance. Experiments often test models, or create environments, that abstract from certain aspects of a situation. If the subjects cannot engage with this abstraction, it becomes difficult to understand their behavior.[3]

Beside professionals and representative samples, the other non-standard subject pools that I will review, namely infrahumans, children, elderly and token economies, are not immune from their own set of issues.

In particular, the point about the limited replicability I made before applies to them as well. Maybe less to animals, to the extent that most universities have animal labs, but nonetheless; the ability to perform experiments with these groups seems much more limited than with the standard subject pool of undergraduate students.

However, some of these samples allow for enhanced control and complete measurement, something that is difficult to achieve elsewhere. For example, in experiments with animals or in token economies, prices can be modified, not simply within the boundaries of a normal experimental period of a few hours, but for a sustained

[3] This is in addition to the more mundane fact that such experiments are often not conducted in the lab, which can limit what one can do and control. This is true also of experiments with representative samples that are sometimes not administered in the presence of an experimenter, but instead online.

period of time: days, weeks, or even months. Similarly, changes in income can happen not only in the context of the experiment, as is the case for standard samples, but also with regard to the total disposable income for the period under study. Finally, all variables can be measured precisely, none of them requiring reliance on self-reports.

Can we formalize how each of these groups is similar or different from the "agent" of our typical model, our subject of interest? Infrahumans have no market interactions, no (or probably very different) socialization, and a different neurological structure (even though they share some commonalities). Children have had no (or few) market interactions, some early socialization, and the same or possibly a different (still developing) neurological structure. Subjects in token economies have had market interactions, the same socialization, and the same or possibly different (malfunctioning) neurological structure. The elderly have had market interactions, the same socialization, and the same or possibly different (decaying) neurological structure. This provides a certain order of distance from the "ideal" subject going from furthest to closest and thus this serves as the organizing criterion to order the sections: infrahumans, children, token economies, elderly, representative sample, professionals.[4]

Writing a review such as this one is challenging for a few reasons. For one thing, papers that have a subject pool in common may have nothing else that connects one to the other. Hence, going from one paper to the next is somewhat disconnected. I have not found a way around this except to try and group papers on related topics together when possible. However, in the discussion, I will come back to papers on a given topic *across* subject pools to tie everything together. The reader interested in a summary organized by choice task can skip directly to that section. Note, however, that many papers reviewed in the previous sections are not mentioned in the discussion as it only considers choice tasks that have been studied using multiple subject pools.

Another difficulty is that, for some of these samples, the number of studies is too large to be covered in its entirety.[5] The solution to this is imperfect and arbitrary, but it is the only one I could find. I will only review papers that came out in economic journals

[4] One should not read too much in this order. I imagine some patients in the token economies are much further from the subjects we want to study than some children are. But it provides a framework. Note also that representative sample and professionals are not seen as ordered by importance. Their relevance depends on the question of interest.

[5] Think of how many experiments with animals must exist on topics relating to economics.

(or working papers that one could expect to come out in those journals). For samples that have fewer studies, such as infrahumans and token economies, I tried to be thorough; for the others, I focused on papers that are either significant in some way or on topics for which there are multiple papers.

Because these samples are unusual, some papers are described in more detail than is typical in a review. In particular, procedures are described in some cases to help the reader better understand the results and how they relate to what is found with other samples. The structure of each section varies slightly, but in all of them I conclude with some observations about methodological issues that are particular to the group under discussion.[6]

II. Infrahumans

The history of economic experiments using infrahumans is about as old as the "rediscovery" of incentivized laboratory economic experiments with humans in the 1960's and 70's.[7] Almost all economic experiments with non-human animals that have been published in economic journals have been conducted by John Kagel and Raymond Battalio who, over the years, have had a multitude of coauthors, the most frequent being Leonard Green, Don MacDonald, and Howard Rachlin. Their experiments used mostly rat and pigeon subjects, although some also used guinea pigs and cats.

As many economists do not even know that there are papers reporting economic experiments conducted on animals, it might even be more puzzling to realize how well published they are. It is even more surprising when considering that the earlier ones were published at a time where laboratory experiments with humans were not common. For instance, of the four papers published in economic journals before 1985, four were in top journals: two were in the *American Economic Review*, one in the *Journal of Political Economy*, and one in the *Quarterly Journal of Economics*. One factor that probably contributed to this success may be difficult to appreciate now: support for the *as if* model. The use of *as if* (optimizing) models of behavior is now so deeply engrained in the

[6] This review is very much in the spirit of Ball and Cech (1996). In addition to the types of groups considered here, they also study robustness within the standard subject pool. That is they consider how results change by educational institution, gender, culture, etc.

[7] Roth (1995) places the beginning of experiments much earlier than what we think of as "modern" experimental economics.

training of economists that it is difficult to appreciate that this was not always the case. In fact, those familiar with work in political science know that rational choice is still debated as an approach in that field. Hence, at the time, showing that infrahumans displayed behavior consistent with the predictions of optimizing agents would have been appealing because it served to support the view that *as if* models can be useful.[8]

At a more fundamental level, experiments with animals allow the researcher to perform manipulations that simply cannot be done with humans. For instance, the quantity of food available to an animal is his wealth / income. When an experimenter changes the quantity of food available to an animal, it is as if that animal's income is changed. These changes can be important both in terms of the size of the income change and the duration of the change. Similar experimental interventions for humans seem problematic and even impossible. Hence, although animals present undeniable limitations, they also offer distinct advantages.

In addition, understanding infrahuman behavior may help to shed light on nature / nurture type questions. If a behavior is observed both in infrahumans and young children, but not in older children and adults; this would seem to suggest that such behavior is eliminated by socialization.

Finally, it is interesting to note that this line of research has served as a conduit for economic ideas into other disciplines: psychology, biology, and neuroscience.[9]

Some aspects of experimental procedures that are common across those experiments are the following.[10] The animals are offered choices through levers or keys between a number of options, namely different kinds of foods and liquids (liquids or food pellets for rats and grains for pigeons), or different time delays to obtain rewards, or different lotteries over varying probabilities and reward amounts. Two basic designs are used: In closed economy experiments, all food intake the animal receives are from within the experiment. In open economy experiments, the food at stake within the experimental trials is in addition to some amount of food and water in their home cages. The amount of

[8] It seems unlikely that animals consciously choose consumption bundles by figuring out the point where their indifference curve is parallel to their budget constraints. In that sense, our models are *as if*, they are not descriptive models of the choice process.
[9] See for instance Rachlin et al. (1980) and Rachlin et al. (1981).
[10] A number of the procedures and equipment for running these experiments were adapted from animal experiments in psychology.

food and water outside of the experiment is sometimes unlimited while in some other cases it is controlled to maintain the animal's weight. The animal practices the task for a period, usually in forced-choice tasks (i.e. the animal is forced to try each lever), then there is a period where the animal explores, and at some point the animal's behavior stabilizes (on average). A stability criterion is used to determine what part of the data will be analyzed and when to change treatments. Typically the number of subjects is small (between two and eight subjects per treatment) and the number of decisions large. Many such experiments use an ABA within-subject design (meaning first control, followed by treatment, and back to control) where the return to the original condition is used to verify that the behavior returns close to its original level in the first A block. The types of choice studied include: commodity choice, labor supply behavior, choices over uncertain / risky outcomes, and intertemporal choices.

Kagel and Battalio produced a sizeable body of work (over 20 papers and a book (Kagel et al. 1995) about infrahuman experiments), with the first published paper appearing in 1975. In that paper (which includes multiple coauthors) they study consumption changes as a reaction to changes in the budget set to determine if those changes are consistent with Slutsky-Hicks demand functions (which we would now think of as testing the General Axiom of Revealed Preferences; GARP). In other words, after an income-compensated price change, does the consumption of the good with the lowered price increase? These questions over basic properties of choice (revealed preferences and the shape of demand functions) were the impetus for much of these coauthors' early experiments.

The design of their experiment, which involved rats as the subjects, is the following: There are a fixed number of lever presses per day (corresponding to the subjects' *income*) to be allocated by the subject across the levers. Two levers, one for each of the two different food and/or liquid options, were available. Changing the total number of lever presses available changes what corresponds to income in a budget set. Either changing the quantity of good per lever press, or the number of lever presses for each unit of good delivered, results in a change in the relative price of the two goods.

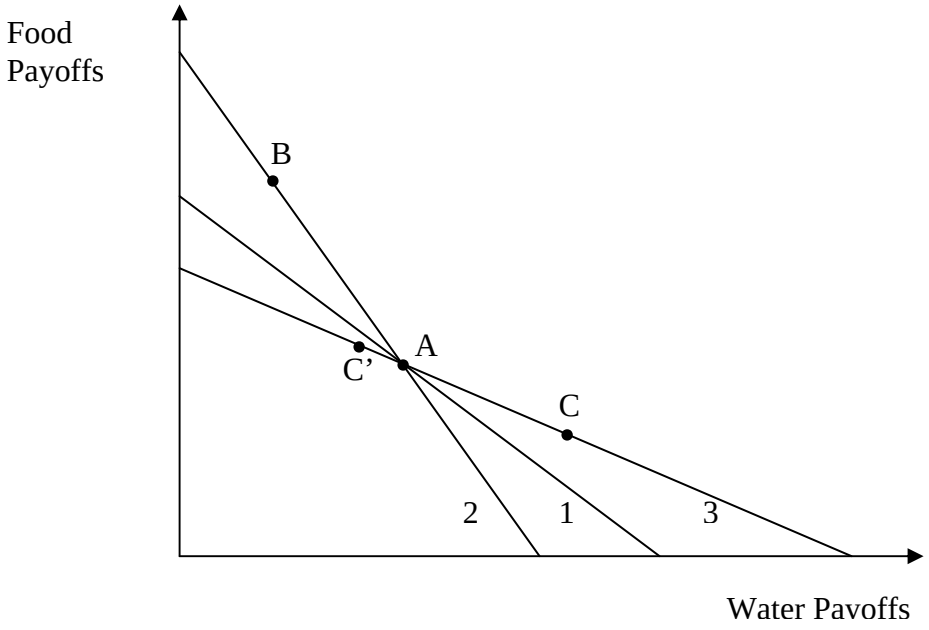
The advantages of using rats for this study are that reporting and recording error of price and consumption are negligible.[11] Income can be precisely controlled when price change, while the commodities in question can be more or less substitutable (e.g., food or water or root beer and Tom Collins mix). Finally, environmental factors can be controlled.

Treatments consisted of changes in the price of the goods while adjusting income to leave the original bundle just affordable. The data is consistent with Slutsky-Hicks demand functions -- the animals consume more of the good that has fallen in price, resulting in downward sloping income-compensated demand curves. Furthermore, the demand is quite responsive to price in the open economy treatments where the commodities employed were inherently more substitutable (e.g., root beer vs Tom Collins mix), while it responds very little to prices in the closed economy treatments where the goods were inherently less substitutable (e.g., food and water).

This question was revisited using pigeons (Battalio et al. 1981) for which they verified the earlier results. That is, subjects consumed a previously unattainable bundle after an income-compensated price change. However, in a second series of studies that involved two price changes, subjects failed to adjust consumption enough to satisfy GARP. Figure 1 gives a representation of this. First, a subject faced budget 1, and consumed A. When the budget is changed to either 2 or 3, consumption adjusts to something like B or C respectively. However, in the second wave of studies, the authors change the budget line to 2 first, and consumption moves to B, but they follow this by a move to budget line 3, and observe a movement in the expected direction, but not quite far enough to satisfy GARP, for example C'. Although the authors do not determine exactly what causes this failure, the result is somewhat suggestive of some form of anchoring.

[11] In typical observational data sets, most variables suffer from some level of measurement error. The same is not true of experimental data, with humans or otherwise. However, when it comes to human experiments, income and consumption outside of the experiment suffers from the same potential problems as survey data.

Figure 1



Clearly, the above results imply downward sloping income-compensated demand curves, and this is explored in Kagel et al. (1981). However in this paper they go further and verify the simple law of demand, namely that (non-compensated) demands are downward sloping, and they do this for both essential and non-essential commodities (food and water versus root beer and Tom Collins mix). Many of these results (and others that will be discussed in this section) were later confirmed in experiments with humans. Having established that basic principles of economic behavior applied to infrahumans, the investigators could sensibly proceed to use these subjects to look for a phenomenon which would be difficult to test for in humans: namely the existence of Giffen goods. Giffen goods provide a stark theoretical example of the importance of distinguishing between income and substitution effects, but are they simply a theoretical construct or can they be observed in practice? Using choices over quinine and root beer (rats prefer root beer to water and water to quinine), Battalio et al. (1991) showed the existence of upward sloping demand for quinine in 50% (3 out of 6) of their subjects. Further, using straight income shifts, it was determined that quinine was strongly inferior for the rats

with upward sloping demand curves, and was either weakly inferior, or a normal good, for those with downward sloping demand. This confirmed that Giffen goods are not simply a theoretical curiosity, and that their existence, in this case at least, follow from the factors the theory postulates.

Beside choices over commodities, the authors also explored choices over consumption and leisure. This was accomplished by a simple change in procedures: now large numbers of lever presses, or key pecks, were required to gain access to a single commodity, with no restrictions on the amount of *labor* supplied. Originally presented in Battalio et al. (1979) but analyzed in more detail in Battalio et al. (1981), these authors show that leisure is a normal good. These results are replicated in Battalio and Kagel (1985) using rats in experiments that excluded a variety of competing hypotheses to explain the animals' behavior.[12] The data on consumption and leisure tradeoffs from pigeons and the data on choices over commodities from rats are analyzed together in Battalio et al. (1987) to understand if such choice behavior in rats and pigeons can be reconciled with utility maximization as opposed to some other type(s) of behavior that mimic utility maximization. The approach is to first estimate demand functions. Then to assess to what extent the estimated parameters correspond to the implications of various models. Specifically they consider six models. Three models that are not utility maximizing models: random "money" deciders and random goods deciders (see Becker 1962), and the matching law, the dominant model of choice at that time with respect to animal behavior (Herrnstein (1961)). They also consider two models that can be conceptualized as utility maximizing models: the minimum distance to a preferred point (quadratic utility functions) and the generalized minimum needs hypothesis (constant elasticity of substitution demand functions). Finally, they test a sixth model, which they refer to as the representative consumer model that can be applied to either of the two utility maximization models. The model that fares best is the minimum needs hypothesis, indicating that their results are consistent with some form of utility maximization. However, the validity of standard aggregation methods for establishing a representative consumer model is soundly rejected.

[12] The variations have to do with possible confounds due to the procedures, for instance they change from an open to a closed economy. Other changes in procedures are to address alternative explanations from reinforcement psychology.

Another important group of papers tackles questions of decision making under uncertainty. It starts with Battalio et al. (1985), which explores risk attitudes in rats and tests for transitivity over choices. They also investigate risk attitudes at varying levels of consumption going from levels resulting in rapid weight loss to levels near satiation. Finally, they test for violations of the independence axiom of expected utility theory (EUT). Rats choose lotteries that give them different quantities of food with certain probabilities. The findings are that rats display risk aversion at all levels of consumption, including where consumption levels are insufficient to maintain weight. Finally, they observe violations of the independence axiom similar to those observed at the time in psychology experiments with humans using non-incentivized, description based choices.

The exploration of the violations of the independence axiom is studied in greater detail in Kagel et al. (1990), with the domain of study changed to focus on losses. The type of violations studied are Allais-type common ratio violations. That is, imagine subjects make two choices. First they choose between A and B, and then between C and D:

Choice 1

A:	x_2 with probability p	B:	x_1 with probability q
	x_3 with probability $1 - p$		x_3 with probability $1 - q$

Choice 2

C:	x_2 with probability rp	D:	x_1 with probability rq
	x_3 with probability $1 - rp$		x_3 with probability $1 - rq$

where $p > q$, $x_3 > x_2 > x_1$, and $0 < r < 1$. A subject that chooses A over B (B over A) and then chooses D over C (C over D), violates EUT, and more specifically, the independence axiom. Some generalizations of EUT can accommodate such choices; in particular the generalizations considered in that paper are the ones that generate indifference curves that fan out (they are not parallel lines in the Marschak-Machina triangle). To allow for losses, defined here as an outcome where less is better, the design is changed by using time delay to payoffs as the random variable. What they find is that although they can reproduce Allais type violations (consistent with the fanning out hypothesis), in other areas of the unit probability triangle they find fanning in, contrary to a leading explanation for common ratio effects at the time (Machina, 1987). They also

show that similar results are obtained in experiments with human subjects. They followed up with further experiments documenting failures of EUT in MacDonald et al. (1991). In that paper they again confirm failure of the independence axiom and observe both fanning out and fanning in of indifference curves. Additionally, they directly test the betweeness axiom (a weakening of independence) and observe choices inconsistent with it. In particular, they observe not only fanning in and out, but also convex indifference curves.

These and other results from the many experiments conducted by these authors are summarized in "Economic Choice Theory" by Kagel et al. (1995). One experiment that stands out, originally published in Kagel and Green (1987), studies time preferences in pigeons. They found preference reversals or dynamic inconsistencies. In particular, they observe a preference for immediate reward over a delayed larger reward, but if the options are both delayed, they observe a preference for the delayed larger reward. One interesting observation is that when these authors were writing, they believed this was one of their first observations of behavior where non-human animals might be different from humans. However, an important literature has now established similar behavior in humans (Frederick et al. (2002)).

To the best of my knowledge, after Battalio and Kagel's last publication of an experiment with non-human subjects in 1991 (with Carl Kogut), 15 years passed before another experiment with infrahumans was published in an economics journal.

In Chen et al. (2006) the authors report two series of experiments. In both of them, capuchin monkeys enter voluntarily in a chamber adjacent to their cage where they can trade tokens for food. Each experiment has multiple sessions, each session involves 12 trials. A monkey is not allowed more than 2 sessions per day. This is an open economy design. In a trial, the monkey (who has a budget of tokens at his disposal) can trade a token with one of the two experimenters on each side of the chamber. Both experimenters present a cue (a certain number of pieces of food), the monkey gives one of the two experimenter a token, and that experimenter gives him a certain number of pieces of food.

In both treatments of experiment 1, the monkeys have a fixed budget and a choice of two fruits (presented to them before they make their choice). The second treatment

differs from the control by changing the relative price of the two fruits and adjusting the budget to make the original bundle affordable. This experiment is similar to Kagel et al. (1975) in that it allows testing if the monkeys adjust their consumption in the direction predicted by GARP. The experiment involved 3 monkeys.

In experiment 2 there are three treatments. Each of them always involve only one fruit, but the two options vary the amount (deterministically or stochastically) as well as what is presented to the monkey before making his choice (either the amount he will receive, less than the amount to be received, or more than the amount he will receive). In the first treatment, one experimenter always presents and gives a single piece of food while the other experimenter presents one piece of food but gives either one or two pieces (each with a 50% probability). In the second treatment, both experimenters provide a 50-50 gamble over one or two pieces of fruit, but they differed in whether they initially displayed one or two pieces of fruit, framing the marginal piece of fruit as a gain or loss. In the third treatment, both experimenters give one piece of fruit with certainty, but differed only in whether they initially displayed one or two pieces of fruit. The experiment involved 5 monkeys.

The results from the first experiment establish that, as in Kagel et al. (1975), the monkeys' behavior is consistent with GARP. It also serves to show that their (novel) procedures, when applied to new and unexplored issues, were not likely to be responsible for the behavior reported. Note, however, that Kagel et al. had more price changes, making it more difficult for the behavior to be consistent with GARP.

The results from the first treatment of the second experiment are that monkeys prefer the option that stochastically dominates the other (in 87% of trials). This is similar to results from earlier experiments on rats using procedures with forced-choice trials.[13]

The second treatment (where both choices result in the same lottery) reveals that the monkeys prefer the option which presents one rather than two pieces of food (in 71% of the trials). Hence their choices exhibit reference dependence in that they prefer the frame in which they make gains rather than losses.

[13] Results vary with rats depending on procedures, going from choosing exclusively the dominant option (in a simple maze with forced-choice) to a very small preference for the dominant option (one of the treatments conducted by Kagel, Battalio, and Green using their type of procedures; for a more complete description see Kagel, Battalio, and Green,1995).

The third treatment finds that 79% of choices are in favor of the option that displays a single piece of fruit (in this case both options yield the same certain reward of one piece). Hence, treatment two and three taken together show that the option that presents only one piece of fruit increases in popularity by eight percentage points when going from a situation where the outcome is a lottery (over one or two pieces of fruits) to a situation where monkeys receive only one piece of fruit. From this they conclude "… this effect [reference dependence] is not confined to risky choices and, when combined with experiment 2, suggests that capuchins are not just reference-dependent but loss-averse", since losses loom larger than gains in both cases.

The first and last papers summarized in this section (Kagel et al. (1975) and Chen et al. (2006)) illustrate two extremes with respect to economic theory. The first experiment established that animal behavior was in line with basic economic principles. The second one suggests the presence in animals of biases that have been documented in humans and which violate basic expected utility theory. This is also consistent with much of the research conducted by Battalio and Kagel. That is to say that experiments on static models of choice between commodities and labor-supply behavior tend to find support for the predictions of standard economic theory. However, when moving to choices over uncertain or risky outcomes and when choosing over time dated goods, animals exhibit biases similar to those of humans, whether it be Allais-type violations, changes over risk preferences between the win and loss domains, or temporal inconsistency in choices. Finally, results indicating context-dependent choices in the spirit of some of the Chen et al. (2006) results have been observed in other species, for instance bees, gray jays (Shafir et al. 2002), and hummingbirds (Bateson et al. 2003).

Methodological Notes

By the very nature of the subject pool, experiments with infrahumans must let the subjects learn about their options through experimentation. On the other hand, most experiments published in economic journals describe the options to subjects such that a subject could theoretically determine what he wants to do without any experience. Some psychologists however use methods similar to those used on non-humans with humans. As discussed in detail in Erev and Haruvy (2014, Chapter xx of this book), experiments using the paradigm of *decisions from experience* often lead to different results than those

based on the paradigm of *decisions from description*, even for some very well-established phenomena. More specifically, contrary to the evidence from the research on prospect theory that subjects overweight events with small probabilities, when using a decision from experience paradigm subjects underweight the probability of rare events. The relevance of this finding for experimental research on infrahumans is evident. When a difference in behavior between infrahumans and humans is observed, care must be taken when attributing the source of the difference to either socialization or cognitive ability. In some cases, the source of the divergence could simply be the different experimental methodologies.

III. Children

Experiments on children have a long history in psychology, at least dating back to early in the last century. In economics, however, their history is much more recent, with the first published paper reporting an experiment with children as subjects appearing in 2000.[14] Much of the early research using children in economics has been conducted by William Harbaugh, Kate Krause and varying coauthors, although others have contributed to this literature.

Just like with non-human animals, if a behavior is observed in young children as well as in adults (the regular subject pool), then it is suggestive of a more robust and universal phenomena, as young children have had much less exposure to culture and to market institutions. Similarly, if a behavior is observed in young children, but not in older children and adults; then it suggests a learned behavior. Understanding when and how humans learn is of interest in itself. However, it also is noteworthy because such behavior is less likely to be universal, as it is mediated by the specific culture and education of the group being studied. Suppose, for instance, that experiments conducted in the United States with the standard subject pool observe a certain behavior, and experiments with American children reveal that this behavior is not present in children of a certain age. This would suggest that similar experiments with the standard subject pool of different countries with different cultures may be worth pursuing.

[14] The paper by Peters, Ünür, Clark, and Schulze existed as a working paper at least as early as 1997, but was not published until 2004.

Also of interest is the fact that children have little to no market experience. Using variation across children in how much market interactions they have (some children enter the labor market early by babysitting, others have a weekly allowance to spend as they wish) may help to understand the role of market institutions in mediating economic behaviors.

Finally, studying children and their parents can help understand the extent of and the process of cultural transmission. Relatedly, it allows us to explore questions such as "where do preferences come from?" (Think about risk or time preference, for instance.) "Are they learned from our parents, and thus malleable, or are some of them transmitted genetically?" "Is there an age at which these preferences stop changing?" Answers to these questions matter for public policy.

Harbaugh and Krausse (2000) study altruism using a typical linear public good game, also known as a voluntary contribution mechanism (VCM) game. Subjects are allocated tokens that they can keep for themselves or invest in a public account which has a known return for all members of the group (the return to the donor of this investment is known as the marginal per capita return or MPCR).[15] Parameters are such that everyone investing everything in the public account is efficient, but keeping tokens in the private account is individually rational. In their study, group size is 6 and the number of repetitions is either fixed at 10 or random (between 4 and 8). They have treatments with a MPCR of 0.50 and 0.33. Participants are 1st through 7th grade students. The experiments are conducted with poker chips that can be exchanged for goods (fancy pencils, small stuffed animals, super balls, etc) at the end of the experiment. Subjects are given 5 tokens before each round and the exchange rate corresponds to approximately 10 cents per token. The results indicate that initial contributions are positive and that they react positively to the MPCR. Furthermore, students who have spent a longer fraction of their life at the school where the study was conducted contributed more. Many other factors considered by the authors were found to have no statistically significant effect: age, gender, number of siblings, single parent household, allowance, TV watching, or church attendance. Unlike in VCM games with the standard college age population,

[15] Group contributions are multiplied by a certain factor, the MPCR is that factor divided by the number of players.

contributions increase slightly over time, except for children 11.5 years and older for whom iterations decrease contributions, as with college age students.

Peters, Ünür, Clark, and Schulze (2004) also study the VCM game. Their main interest is to test the notion that, if parents are altruistic, then it might be in the self-interest of children to maximize family income even if they are not altruists themselves (Becker's Rotten Kid Theorem). In a VCM game this would predict that children would contribute to the public good when playing with their family (but not with strangers) because they expect returns later on. The adults' range in age between 34 and 50 (average 42) while the children are between 9 and 16 (average 11). The game is repeated 24 periods in blocks of 8 periods where participants play first with strangers, then with their family, and then strangers again; or family, strangers, family. Families vary in size (3 or 4 members) and composition. Some key findings are that: 1) Children give more to family members than to strangers. 2) Parents give more to family members than to strangers. 3) Parents give more than children in both conditions. 4) Contributions tend to decrease over time in every condition. To better understand the reasons why parents give to both their children and other children, they conducted an additional (strangers) treatment in which parents are grouped with other parents who they would not be familiar with. There is no statistical difference between contributions in the family condition and the strangers' condition (where they play with other parents). Note, however, that the authors compare all periods and do not report comparisons for the last periods. Looking at their figures, the drop in contributions is more pronounced for strangers than families, and thus they may be different by the end.

Taking the data of both studies together suggests that the standard results from VCM experiments, which is that contributions decrease over repetitions, might only appear in the early teens. Peters et al. (2004) do not find an effect of age on contributions, but their children are older than those used by Harbaugh and Krausse (2000). Other results are in line with findings using undergraduate students: contribution levels react positively to the MPCR. Both studies also suggest that close attachment to a group (either a family or a school in which one has received most of their education) leads to higher contributions.

Other papers have investigated other-regarding preferences in children. Harbaugh, Krausse, Liday Jr., and Vesterlund (2002) study trust games (also known as investment games). The general structure of a trust game is the following: The *A* player has some tokens, dollars, or other quantity to start with. He has a choice to send something to player *B*. What is sent is multiplied by a known factor. The *B* player can then send something back to player *A*. The final payoff for player *A* is equal to what he kept, plus what he received from player *B*. The payoff for player *B* is equal to what he kept. Player *A* is sometimes referred to as the truster and player *B* the trustee (or first and second sender respectively). The specific version of that game implemented in this project gave four tokens to player *A*, and what was passed to player *B* was multiplied by three. They used the strategy method, meaning that *B* players indicated what they would do for any of the five possible amounts *A* might send them. They used children from third, sixth, ninth, and twelfth grade. They asked the *A* player to make multiple transfers, one to a child of each grade and one to an adult. The adult's decisions were mimicked by averaging the decisions of subjects in previous experiments. *B* players were asked for their returns separately for each possible grade of player *A*. For third graders, each token was worth approximately 25 cents at a portable toy store brought to school. Other grades were paid 25 cents per token. The findings indicate very little variations in what *A* players send to *B* as a function of *B*'s grade (age). On average 1.32 tokens (out of the 4) are passed. The amount *A* players send decreases as their score on a survey measure of trust increases (and thus suggests that this survey measure may be misleading), the birth order of the child has a positive impact, and the relative height of the child also increases the amount sent. Note that the age is found to have no effect. As for *B* players, the main factor affecting what they return is what they were sent. Nothing else beside the *A* player's grade has an impact, and that is a small positive one. For example, the predicted difference between what is sent back to a third grader and a sixth grader is only one tenth of a token.

Other-regarding preferences are investigated through individual decisions about allocations of payoffs to a group in Sutter et al. (2010). Using Austrian children in grades three, five, seven, nine, and 11, the experiment consists of selecting one of three distributions of payoffs for a group of three players. Subjects are matched in groups of

three, with randomly assigned positions one, two, and three; with the preference of position two implemented (thus two receives the middle income). This is done eight times, each time with a different set of options. The authors use those choices to categorize subjects according to certain types of other regarding preferences. Incentives were varied for younger and older kids but were all monetary. The findings are that some choices do not vary with age while others do. These patterns are different for boys and girls. More specifically, choices that the authors categorize as corresponding to selfish preferences and ERC type preferences (Bolton and Ockenfels (2000)) are similar for both genders. While choices corresponding to Fehr and Schmidt (1999) type preferences are decreasing with age for both genders. When it comes to efficiency, it is fairly constant across ages for girls while it increases for boys and MaxMin type preferences increase with age for girls, while it is mostly constant for boys. One caveat is that preferences are not uniquely identified in all choices. In particular, selfishness is only uniquely identified in 2 choices, while efficiency is uniquely identified in all choices. This can lead to problems. For instance, imagine that subjects prefer efficiency at all ages, but sometimes make mistakes. However, they make fewer mistakes as they become older. Then it is possible for this to result in a decreasing fraction of subjects characterized as selfish and an increasing fraction of efficiency types, when, in fact, preferences are stable.

Another area of research with ties to questions about other-regarding preferences is bargaining. Harbaugh, Krause, and Liday (2003) study the behavior of children in second, fourth, fifth, ninth, and twelfth grades in a dictator game and an ultimatum game.[16] In the ultimatum game, a proposer has control of a sum of money, he offers a division of that money to a responder who can accept, in which case they each get their part of the proposed division; or reject, in which case they both get nothing. The dictator game is similar but eliminates the possibility for the second mover to accept or reject (the dictator game is an individual decision problem). For both games, subjects played in both roles (something which is not standard in adult experiments). They had 10 tokens to play with (per game), these tokens were exchanged for goods at a rate of about 25 cents (per token) for the youngest children, they were exchanged for money (25 cents) for the ninth

[16] Bettinger and Slonim (2006) study the choices of children in a dictator game and analyze the impact of a natural experiment on their behavior.

graders, and 50 cents for the twelfth graders. They find that the tokens offered increase substantially in both the dictator game and the ultimatum game going from grade 2 to grade 4. This increase is fourfold in the dictator game, going from 3.5% to 14% of the endowment, while for the ultimatum game it goes from 35% to 41%. Rejection rates are relatively constant at about 10%, despite the differences in offers. Hence, younger children accept lower offers than older children. At all ages, average offers are substantially higher in the ultimatum game than in the dictator game, indicating that even young children react to the strategic nature of the games.[17]

Harbaugh et al. (2001) study choices in seven and 11-year-old children and college undergraduates to see to what extent they respect GARP. The subjects had to select from 11 different choice sets. Each choice set consisted of a finite number of alternatives (between three and seven bundles). The goods were small bags of chips and boxes of fruit juice. Once all the choices were made, one of the 11 choices was randomly selected to be consumed. The subjects were shown all 11 choices three times and they had the opportunity to change their choice each time. The results indicate that some violations are present at all ages but much less than if choices were random, with these violations decreasing with age. Random choices would yield 81% of violations, whereas second graders have 39% of violations, 6th graders exhibit 19% of violations, and undergrads have 18% of violations. However, the severity of violations as measured by Afriat's index does not change much with age.[18] Random choices in this study would result in an Afriat index of 0.65. What they found is an index of 0.93 for second graders, 0.96 for 6th graders, and 0.94 for undergraduates and these numbers are not statistically different.

List and Millimet (2008) use a task similar to Harbaugh et al. (2001) (they also have 11 choices over chips and boxes of juice) to explore the impact of market activity on GARP violations.[19] Their experiments were conducted at malls that also host sportscard shows. There they recruit youths ages six to 17 (on average about 11) to participate in their experiment. They divide them in three groups: two treatments, one where the

[17] Harbaugh et al. (2007) report that children from age eight to 18 make offers close to optimal given responder behavior.
[18] The Afriat index measures of how much the budget constraint would need to be relaxed to accommodate the choices being consistent with GARP - no violations results in an index of 1.
[19] In addition, the paper reports the results of a market experiment with children not summarized here.

experiment ends after the GARP task, and one where subjects are given a gift of sports cards they could trade at the show after doing the GARP task. Subjects who had never previously attended a sportscard show were randomly assigned to these two treatments. The third group consists of subjects who had prior experience at sportscard shows (they were placed in the no gift treatment). Seven months later, these subjects were invited back to do the GARP task again: the two different dates are referred to as round one and round two. Attrition between the two dates was important with only 420 subjects of the original 819 coming back in round two. They report that in round one, subjects who had not previously been to sportscard shows displayed about four violations (out of 11) whereas subjects who had been to such shows only displayed about two mistakes. Seven months later, the violations fell in all groups (an average reduction of between 0.3 to 1.0 choices), but they fell the most amongst subjects assigned to the gift treatment who then decided to participate in the sportscard show. The authors use econometric methods to tackle the various challenges of this kind of data (for instance, the mass of subjects with no violation) and to control for confounding factors. Taking the data at face value, they find that market experience decreases GARP violations. However, once they control for endogeneity, the effect disappears. On the other hand, age is found to have an impact in most specifications (older children make fewer mistakes).

More specifically, distinguishing the effect of market experience amongst these treatments is not easy as there are multiple levels of endogeneity: who was already going to sportscard shows is not exogenous; when assigned to the no-gift treatment, afterwards actually not attending the show is not exogneous; when assigned to the gift treatment, afterwards actually attending the show is not exogneous; and coming back for round two of the experiment is not exogneous. Hence, the paper reports multiple regressions, some performed on a subset that excludes subjects with prior experience. In addition, some specifications exclude the subjects whose behavior after the experiment did not correspond to the desired treatment (i.e. subjects in the gift treatment who did not go to the show or subjects in the no-gift treatment who did). Note that this does not eliminate endogeneity, but does reduce the problem. They also perform an estimation where they preserve these subjects but instrument for going to the show once in the experiment. Going from the full sample to the restricted sample reduces the magnitude of the effect of

market experience by half (from about one to around 0.4 -- that is 0.4 less mistakes out of eleven choices). Once they instrument for participation at the show the estimate drops to 0.15 fewer mistakes out of eleven choices, but the effect is no longer significant. Note, on the other hand, that the effect of age is significant (in all but one specification) and the magnitude of its impact is always at about 0.5 for the full or restricted sample, with or without IV (this means 0.5 less mistakes per year).

Some of the results of List and Millimet (2008) do not line up with those of Harbaugh et al. (2001). In particular, Harbaugh et al. (2001) report on average 4.3 violations (out of eleven) for seven years old and 2.1 violations for 11 years old. The List and Millimet (2008) subjects who are on average 11 make about 4 mistakes and the ones who have prior experience at sportscard shows make 1.9 mistakes. In other words, the behavior of the eleven year olds of Harbaugh et al. (2001) is similar to subjects with market experience in List and Millimet (2008). Similarly, the youngest subjects in Harbaugh et al. (2001) behave much like older subjects in List and Millimet (2008). These differences could be due to differences in procedures or subject pools. Together, however, they both support the view that GARP violations decrease with age (at least between seven and eleven). The evidence with respect to the impact of market experience is less clear.

In another paper, Haurbaugh, Krausse, and Vesterlund (2001) investigate the presence of the endowment effect in children attending Kindergarten, third grade, and fifth grade, as well as amongst undergraduates. Subjects are given one good (the endowment), and then offered to trade it for a different one. The choice is repeated for four pairs of goods (three in the case of undergraduates). The authors measure the endowment effect as the difference between the probability that the subject will choose good *A* when he is endowed with good *A* versus the probability that he will choose *A* when he is endowed with a different good. The endowment boost is defined as the average, between the two goods, of the probability that a subject chose a good when they were endowed with it as opposed to when they were not. The average endowment boost across pairs of goods is 3.1 for kindergartners, 1.5 for third graders, 3.9 for fifth graders, and 3.5 for undergraduates. But there are no statistical differences in the endowment boost across ages.

Finally, Bettinger and Slonim (2007) study inter-temporal choices amongst children ages five to 16. They also relate the children's behavior to information about their parents. The children are offered compensation (in the form of Toys-R-Us gift certificates) at two sets of dates: the first set being immediately or two months in the future, and the second set being 2 months or four months later. A key result is that more children are consistent with hyperbolic discounting than not. They also find that patience increases with age and boys are less patient than girls. About a quarter of children make choices inconsistent with any rational model of choice, and this fraction decreases with age. Finally, family income, the parent's education or the parent's patience (measured as for the children but using money), do not significantly correlate with the children's patience.

Methodological Notes

Two issues that arise with children are 1) how to incentivize them and 2) how to explain the task? With respect to incentivizing , for children of a certain age, money is appropriate and many of the studies with children simply use money, but for younger children such as kindergartners, it may not be ideal. The approach used by Harbaugh, Krausse and coauthors is to pay children with tokens that can be used to buy toys in a moveable store they bring with them and show to the kids before doing the experiment.

Explaining procedures to children forces the investigators to be extremely clear, which is always good practice. However, the need for clarity and simplicity sometimes makes it impossible (or at least impractical) to use certain methods, such as the Becker-DeGroot-Marschak (BDM) mechanism (Becker et al. 1964) to elicit valuations. Hence, just as with animal experiments, certain comparisons across the standard subjects and children are confounded with methodological differences. To the extent that results show similarities across the samples, this may not be an important concern, but in cases of differences, it is certainly worth thinking about.

IV. Token Economies

Token economies were developed by psychologists as a mechanism to modify the behavior of institutionalized individuals. They are mostly closed systems in the sense that individuals who are part of token economies spend the majority of their time in that

environment and their earnings and consumption also occur for the most part in the system. Typically, individuals in token economies earn tokens for work such as making beds, cleaning bathrooms, and simple factory jobs. They can spend their earnings on food, cigarettes, movies, clothing, etc. However, as will be seen below, token economies are not limited to institutionalized individuals.

Experiments in token economies that address economic questions started with Ayllon and Azrin (1965). They changed the relative wage of different tasks to determine if wages can be used to induce individuals to change to their less preferred work. They found that indeed they could influence job selection. More studies followed in psychology (see Tarr 1976).

The first mention of token economies in an economics journal is in a communication published in 1972 in the *Journal of Political Economy* authored by John Kagel. He makes an argument for why token economies might be useful in testing economic theories. John Kagel and Raymond Battalio were students of Robert Basmann, an econometrician at Purdue at the time. Basmann wanted data to be more closely connected to the primitives of the models. For instance, having choices of multiple subjects over time allowed tests of GARP without making the strong aggregation assumptions that are necessary if one only has cross-sectional data. This desire for data that corresponded to the primitives of the model led Kagel and Battalio first to token economies, effectively very early field experiments, and then, via their psychologist collaborators, to infrahuman experiments.

Panel data are now much more common and some of the advantages that led Kagel and Battalio to use token economies are more easily available nowadays using the standard subject pool of undergraduate students. Nonetheless, token economies present a few advantages over the standard subject pool. The most important one is that, unlike in a standard experiment, the entire economic eco-system is observed. In a standard experiment, much of the income and consumption of subjects happens outside of the laboratory. The same is not true in a token economy, and thus a token economy provides more control. On the other hand, this control is limited, as it cannot interfere with the primary purpose of the token economy, which is to reinforce “good” behavior.

A token economy also provides precise measurement of variables of interest. Although this is in part true of standard experiments, it is not always so. In particular, behavior that may depend on total income (not simply experimental income) can only rely on reported income in a standard experiment. In a token economy, the tokens one owns are his income.

The first actual study using data from a token economy to appear in an economics journal is Battalio et al. (1973). They used data from the female ward for chronic psychotics at the Central Islip State Hospital in New York. The experimental variation was to change the relative prices for bundles of goods. That is, each good was placed in one of three groups, and the prices of goods in that group were sometimes doubled and sometimes halved. The sequence of changes followed the standard ABA design. The experiment took place over a period of seven weeks and relative prices were the same for 1 to 2 weeks (although this was not known to the subjects). Patients were also unaware that they were taking part in an experiment. The experiment focuses on whether individual consumption patterns are consistent with revealed preference theory (GARP). Furthermore, they highlight the role of recording mistakes in limiting the ability to test GARP. They show that even small reporting errors can lead to the incorrect rejection of the model. In their data, they have two independent measurements of consumption (the records taken at the time of sale, and the total tokens spent per patient in a week – tokens had the patient's name stamped on them), and thus they can evaluate the extent of measurement error (which varied on average by 3.6%; Battalio et al. (1977)). Although they use the sales record for the first test, in case of a rejection, they then look for any possible mis-measurement consistent with the week's total that would render the observations consistent with GARP. Only when both tests fail do they consider the data to reject the theory. What they find is that for 19 of the 38 subjects, the data satisfies GARP for all weeks. For most of the remaining 19 subjects, the contradiction occurs in a single pair of weeks. In fact for 17 of these 19 subjects, there exists an allocation of the token difference between the two measures such that the data from all weeks is consistent with the model. Hence, behavior of only 2 out of the 38 subjects clearly rejects the model.

Battalio, Kagel and coauthors studied the Islip patients in multiple additional papers (Battalio et al. 1974, Basmann et al. 1976, Kagel et al. 1977) and their data was revisited in Cox (1997). These papers also investigated varying aspects of demand behavior, except for Kagel et al. (1977), which focused on labor supply decisions.

Another token economy studied consisted of three groups of volunteer subjects in an experiment on the impact of cannabis consumption on productivity. The subjects were paid to produce woolen belts on small portable hand looms at the Addiction Research Foundation in Ontario, Canada, with cannabis freely available, along with required smoking at a fixed time each day.[20] This Cannabis Economy was studied on its own in two papers: Battalio et al. (1978) studying the distribution of earnings and Kagel et al. (1980) which examine the impact of marijuana consumption on productivity.

The Islip data was also used in conjunction with the Cannabis Economy data in Battalio et al. (1977). In that paper, they use these two token economies to determine if, in an economy where the sole sources of variation in earnings are due to variations in ability and effort, the earnings are more compressed than with typical field data. That is, a host of factors could account for differences in labor income in field settings such as chance, differences in cost of training across occupations, differences due to market failures such as inability to borrow the funds to finance training, and others. Some contend that ability and work-leisure tradeoffs do not vary enough across individuals to explain the extent of variations observed in income. In the token economies, the work is simple enough that no training is necessary; everyone can work if they want to, and they can work as much as they want. Hence, the sole sources of variation in earnings in those token economies are ability and leisure decisions. What they show is that there are important variations in income in both cases. For instance, in both the Islip (25 subjects over 5 weeks) and Cannabis (56 subjects over 3 months) economies the maximum income is 10 times as large as the lowest. However, there is slightly more compression in the Cannabis economy where the age distribution, and subject characteristics, were much

[20] The policy question motivating this experiment was the potential negative effect of legalizing Cannabis on worker incentives. In this respect, one of the most interesting side effects reported is that in one session, subjects went on strike *against* having to smoke so much pot as it interfered with their earnings from belt making, which was theirs on leaving the experiment.

more compressed.[21] In fact, the Gini coefficients for both token economies fall within the range of multiple estimates from the US, India, and the United Kingdom. Other summary statistics also confirm that variations in the token economies are similar to variations in earnings in the US. Although this by no means proves that ability and work-leisure tradeoffs explain the variation in earnings at the national level, it does show that those two factors by themselves can generate such variation and that in an environment without nepotism, discrimination, or other barriers to being able to work, earnings can vary greatly across individuals.

The two papers described are illustrative of the work that has been performed using token economies, namely that economic behavior in those environments is consistent with evidence obtained elsewhere and with the predictions of economic theory. For instance Varian (2006) in his review of the history of revealed preferences studies reports that the evidence from these experiments is consistent with aggregate time-series data. The experimental work with token economies also manages to venture in areas that cannot be tackled with the standard subject pool such as the distribution of income.

Methodological Notes

Although token economies provide an ongoing economic system to study with much precision and completeness (it is one of the rare occurrence of data with humans where prices can be manipulated and both consumption and income is known), they also impose technical limitations. In particular, many experimental variations of potential interest would be unacceptable. Token economies have largely gone out of fashion in psychology, and an experiment like the Cannabis economy is quite expensive and time consuming, so that it would be hard to get a grant these days to study strictly economic issues in such a setting.

V. Elderly

With the population of developed countries aging and since older individuals control a substantial amount of wealth, the importance of research on the elderly can only grow. This group, however, does not present many advantages as a group to study. It is

[21] In the Cannabis economy subjects were between 20 and 28 years old, as opposed to 19 to 64 years old in Islip.

not convenient (one needs to take the laboratory to them), interacting via computers is still difficult for many elderly, and for some of them, abstract instructions (and even written instructions) can be difficult to follow. One advantage of studying this group is that it allows us to gain an understanding of specific brain functions on behavior since aging does not affect all brain functions equally. However, as I pointed out for the other groups: if a behavior is observed in the standard subject pool and in the elderly, this is further evidence of a robust phenomenon.

The earliest laboratory study of older adults that I could identify in economics is a 2005 publication by Kovalchik, Camerer, Grether, Plott, and Allman. They study two groups, 51 college students aged between 18 to 26, and 50 elderly aged between 70 and 95. The group of elderly subjects are neurologically healthy (they served as a control group for Alzheimer research). The experiment addresses four areas: overconfidence, risk preference, the endowment effect, and strategic thinking. More specifically, in the first task subjects answer 20 trivia questions, after which they are asked their assessment of the percentage they got right (50, 60, 70…,100). This is followed by two gambling tasks in which subjects pick a fixed number of cards from either of two decks, decks A and B. In the first task, they pick 50 cards, while in the second task, they pick six cards. One deck has a negative average payoff (deck A) while the other has a positive average payoff (deck B) and the deck with the positive average payoff has a smaller variance. In the first case, subjects know nothing about the two decks in advance. Hence, there is no optimal strategy, but in neuroscience, a choice of A is seen as a mistake, as they think subjects should learn to pick B. In the second case, subjects are shown the 10 cards in both decks before the cards are shuffled. In a third task, the investigators assign each subject to be either a buyer or seller, and then they elicit the willingness to pay or willingness to accept payment for a mug. To finish, they play a beauty contest game in groups of nine. They select a number between 0 and 100. The subject that states the number closest to $2/3^{rd}$ of the average of the nine numbers wins $20.

Overall, the results indicate more similarities than differences as a function of age:

1. In the beauty contest, both groups show a first cluster of data around 33 and a second cluster around 22.

2. With respect to the endowment effect, they observe no differences in willingness to pay and willingness to accept in either group. Note that the procedures used differ from the ones that commonly find a gap between these two values.
3. Confidence is also similar in that "both groups of subjects display overconfidence at some levels, and neither group shows underconfidence at any level." (p. 82) There is a difference however in that older subjects are more accurate.
4. Lottery choices are also very similar across the two groups. In the first task (unknown probabilities) choices evolve toward mostly selecting B for both groups. In the second task (known probabilities), both groups choose from each deck with similar probabilities, and, surprisingly, display a slight preference for deck A.

There are two other studies of risk preferences in the elderly. Kume and Susuki (2012) compare the behavior of 31 subjects between 65 and 76 years old (from an employment agency in Osaka) to 32 subjects between 25 and 65 (44% are in their 40s) who were employed at the time of the study. The experiment consists of eliciting the willingness to accept for lotteries of a probability p of winning a prize or (1-p) of getting nothing. The probability is determined randomly for each of the 20 decisions, and the willingness to accept is determined using BDM procedures. The authors report a difference between the two groups; but their analysis ignores statistical significance, making the claims difficult to evaluate. Further, the two groups are faced with different lotteries since the p is drawn randomly, which also makes it difficult to compare the two groups. Using the summary statistics they report, the following can be said: 1) Both groups are on average risk averse, as the willingness to accept is lower than the expected value. 2) The stated willingness to accept is lower for the older group and this difference is statistically significant at the 1% level.[22] Hence, in this data set, the older group seems more risk averse than the younger group.

Charness and Villeval (2009) also test for risk preference among older adults. They study junior workers (48, average age 25) and senior workers (39, average age 54) employed in two large private manufacturing companies in France. They also study

[22] I computed this using the reported summary statistics, but as a result, I cannot account for the fact that we have repeated observations for subjects. If there is positive within subject correlation (due to risk attitudes for instance), the standard deviation reported would be too small and the p-value reported too low.

students (37, average age 21) from schools around Lyon and retirees (35, average age 66) recruited through local associations and one municipality. The task consists of deciding how much of 100 points to invest in a risky asset. The investment fails with probability 50%, in which case the points invested are lost. If it succeeds, then the return is 2.5 times the investment. At the end of the experiment, tokens are converted to Euros at a rate of 40 for 1. A risk neutral subject should invest everything. The results are that: 1) Average behavior in all four groups exhibits some level of risk aversion. 2) There is no statistically significant difference in behavior across groups (average investments range from 50 for students to 59 for the working seniors). If anything, the average moves in the direction of older subjects being less risk averse.

Taking the three studies together, it seems that the stereotype that older adults are more risk averse finds little support. In particular, in the one study that finds evidence in that direction (Kume and Suzuki 2012), the older subjects are unemployed and looking for work, whereas the younger subjects are actually not that much younger and employed. This suggests that their older subjects may be in more immediate need of money.

The Charness and Villeval (2009) experiment, however, is mainly aimed at studying cooperation and competition. First, the subjects participate in 17 VCM games in groups of 3 with a MPCR of 0.5. There are 2 blocks of 8, where either the composition of workers (young or old) is known or not known. They conduct sessions in both orders (known then unknown and vice versa). This is followed by a 17^{th} period where subjects first choose the composition of their group, and then play the game. They also conducted a real effort task where subjects solved anagrams for four minutes. Subjects, who were paired, had first to select if they wanted to be paid for the number of anagrams they solved (18 points per anagram) or on the basis of relative performance (30 points per anagram for the winner and 6 points per anagram for the loser). Finally they elicited beliefs about a subject's own performance and about the average performance of juniors and seniors. In the VCM game they find that retirees and seniors contribute more than junior workers, and junior workers contribute more than students. They also find that team composition (and its knowledge) affects contributions, namely heterogeneous groups contribute more. However, in both blocks, contributions trend downward in all

groups. This raises the question of whether they may all be trending toward the same (low) contribution level, simply getting there at different speeds. When it comes to competition, behavior is more similar across groups. Subjects in all groups provided more effort (solving anagrams) when they are in a competition rather than when they are paid for absolute performance. Groups do not differ in productivity if they are in a competition. However, retirees are slightly less inclined to choose the tournament than students.

Holm and Nystedt (2005) perform an experiment over the mail, with names they obtained from a public database in Sweden, where they invite 120 subjects, half of which are 70 and the other half are 20 (in the end 81 subjects participate). The game is a trust game where player *A* can send any part of the SEK 100 they received to a player *B*, the amount sent is multiplied by three. Players also received a flat payment for participating. The main result is an absence of difference between the two groups. They find no statistical difference in the amount sent between the younger and older subjects. They also find no difference in amounts returned, with the mode at one-third of what *A* sent. However, there is more dispersion in amounts returned in the older group.

Besedeš, Deck, Sarangi, and Shor (2012) investigate binary lottery choices using 127 subjects in an online study. The subjects are part of a demographically diverse database, maintained at Vanderbilt University, of individuals interested in participating in online experiments. They recruited 35 subjects below 41, 45 subjects above 60, and 47 subjects in the middle group 41 to 60. The task is the following: A card will be drawn from a deck. The number (probabilities) of each card present in the deck is known (the number of cards is varied across rounds and each card has a different probability of being drawn). Subjects must pick a bet from a pre-specified set of options (the number of options is varied). Each option specifies which card would pay. If the subject selected an option that pays-out for the card that is drawn, he receives $1, otherwise he receives nothing. Each subject performs 8 rounds. This problem is one of selecting the option for which the sum of the probabilities where it pays is highest amongst the available options. In other words, in this problem, the prize is fixed, and subjects must select the portfolio that pays in the most states of the world, so that an expected utility maximizer would select the option that maximizes the probability that a winning card will be selected,

independent of his specific utility function. Overall, subjects selected the optimal option 40% of the time, with the frequency of optimal choices decreasing with the number of options, suggesting that complexity decreases performance. On the other hand, there is no significant effect with respect to changing the number of cards over which options are specified. The key result is that performance declines with age: adults over 60 select the optimal choice in 32% of the problems, while this number is 52% for those below 40 (this difference is statistically significant at the 1% level). This represents a 6 percentage point change in efficiency from 84% to 90%.[23] The small change in efficiency reflects the fact that the cost of mistakes is relatively low in this experiment. They also find that the impact of aging is much less important than that of having a graduate degree. Both in terms of optimality and efficiency, the regressions suggest that not having a graduate degree reduces performance as much as aging by about 40 years.

It is difficult to draw overall conclusions about aging given the small number of studies using older adults and the mixed results: some studies find differences while others don't; there are no systematic patterns yet. This is probably because of important differences in the group of older subjects recruited. The unemployed elderly looking for work, the elderly registered as a control group for studies on Alzheimer's disease, older adults registered to participate in online experiments: these are all potentially very different types of older adults. Furthermore, these subjects are probably very different from a representative sample of the elderly. Hence, to the extent that this type of research is interested in the impact of declining functions in the aging population, a better understanding of variations in behavior within the elderly population would be useful. However, it seems fair to say that some of the stereotypes about aging are not supported by the data so far.

Methodological Notes

Since aging might affect males and females differentially, this might be an area of research where keeping track of gender is particularly important when comparing groups of different ages. Moreover, unlike typical lab experiments where assignment to treatments is done randomly from the subject population, and thus gender is on average

[23] Efficiency is computed as the ratio of the expected payoffs of the actual choices to the expected payoffs of the optimal choices.

represented in the same proportions across treatments, it may be that the group of older adults under study has a very different ratio of males to females than amongst younger adults as women live longer than men.

With older subjects, it is often inconvenient to bring them to a laboratory. Hence, many of the experiments with older adults were performed *remotely*, that is either online or through the mail. Although it is still unclear to what extent results from experiments in the laboratory versus online differ, this could potentially be a concern. As a practical example, in laboratory experiments it is common practice to read instructions aloud so that subjects know that other subjects were presented with the same instructions (knowledge of common information). When moving to experiments online or through the mail, this possibility is lost.

VI. Highly Demographically Varied (Representative) Sample

Recently, some authors have attempted to perform experiments on a representative sample of a given population. This has, for the most part, occurred within already running surveys. This section will review those papers as well as papers which, although they do not have a representative sample, have a highly demographically varied sample.[24] Discussing the two together makes sense because the goal is often to identify the impact of demographics (or other variables) on behavior, something that can be done even if the sample under study is not itself representative.

Three studies using representative samples investigate the trust game. Fehr et al. (2003) use surveys of a representative sample of the adult German population living in private housing. They interviewed 442 individuals, 429 of whom accepted to participate. In their version of the trust game, *A* players have a finite set of amounts to choose from, and the amount received by the recipient is doubled. Subjects were paid by mail after the answers of players *A* and *B* were matched (instead of using the strategy method for *B* subjects, they randomly give an amount to *B* based on the probability distribution with which each amount was chosen in a pilot experiment). The second study, Bellemare and Kröger (2007), recruits through the CentERdata which surveys 2000 households

[24] Note also that even studies that aim to have a representative sample do not necessarily end up with a representative sample.

representative of the Dutch population. As in the Fehr et al. (2003) study, the amount sent by *A* from a finite set is doubled, but unlike them, they use the strategy method for *B*'s answer. The third study, Falk et al. (2013), uses 1,001 subjects representative of the population of the city of Zurich. Their implementation of the game is similar to that of Bellemare and Kröger (2007) except that the amount sent by the first mover is tripled and they conduct the experiment via mail correspondence.[25]

All three studies find that very few factors are significant determinants of behavior, with only two factors consistently impacting behavior: age and the beliefs of player *A*. Not surprisingly, the beliefs of player *A* regarding what player B will send back correlate positively to the amount sent. With respect to age, although they use different regression specifications, some results overlap. For amount sent, both Bellemare and Kröger (2007) and Falk et al. (2013) use age and age squared as regressors and find a concave relation. Fehr et al. (2003) use dummies for blocks of ages and find that adults over 65 give less than those between 35 and 55. Hence, all three studies find that the elderly send less than adults in their midlife. In addition, the specifications of Bellemare and Kröger (2007) and Falk et al. (2013) imply that the amount sent increase earlier in life, but Fehr et al. (2003) does not replicate this finding. Results for amount returned are less homogeneous: Fehr et al. (2003) report a positive relation to age, Bellemare and Kröger (2007) find a convex relation to age, and Falk et al. (2013) do not find that age has a significant impact on the amount sent back.

In addition, Bellemare and Kröger (2007) and Falk et al. (2013) also study a standard subject pool of undergraduates (at the University of Tilburg and at the University of Zurich respectively). Bellemare and Kröger (2007) find that the raw data are statistically different across the two samples. However, the effect is no longer significant once they control for demographics and beliefs. Falk et al. (2013), on the other hand, find that the first mover data is not different across the two samples, with or without demographic controls. The behavior of second movers, however, is different for

[25] Falk et al. (2013) also analyze the volunteer artifact by studying donations to a social fund at the University of Zurich by all students who register to the university for the first time over a period of six years (16,666 students). Students are asked for these donations at the beginning of each semester, and it is used to support foreign students and provide loans for the needy. They find that, although subjects who volunteer to participate in experiments (1,783 of those students) differ from those who do not in terms of their demographics, their donations to that charity are no different. This is true with or without controlling for demographics.

students and non-students (students return less at all amounts, but the slope of the relation is the same). Like Bellemare and Kröger (2007), this is no longer the case once they control for demographic variables.

It is not clear how to reconcile the observation that age affects behavior in these three studies with the result of Holm and Nystedt (2005) who find no difference in behavior in the trust game between 20 year olds and 70 year olds (see the previous section). It could be that there are differences between the Dutch, Swedish, and Swiss populations. Alternatively it could be that because the relations between age and behavior are non-linear, 20 and 70 year olds end up being the same.

All three studies also report other demographic factors, beside age, as significant, but those that are significant in one study are not in another and vice versa. This suggests either false positives, or it could be the result of variations in procedures or in econometric specifications. However, it could also indicate cross-country differences.

Another paper using the CentERdata is by Bellemare et al. (2008). In that paper they investigate behavior in discrete ultimatum and dictator games. Their game is modified such that no equal split option is available (the closest offers being 450-550 or 550-450), and the strategy method is used for the responder. Beliefs of proposers were also elicited, although those were not incentivized. A related game, the three-player ultimatum game, is studied by Güth et al. (2007) in a newspaper experiment. The experiment was conducted as a contest in *Die Zeit*, a well-known news magazine in Germany (1.03 million readers per issue). In total, 25 participants would be randomly selected with each able to win up to DM 1,000 per person. The three-player ultimatum game is one where a proposer can select a division of a fixed pie (DM 1,200) between three players, in this particular case according to a finite set of options. One of the other two players is selected to accept or reject the proposal, while the third (dummy) player accepts whatever she has been allocated as in the dictator game. If the proposal is accepted, it determines the final payoffs, and if it is rejected, they all receive nothing. The game is played using the strategy method with all three players choosing in all three roles. They were also asked to predict the most common behavior in the role of proposer and responder. Given the specific offers available, (1000, 100, 100) ordered with the proposer first, is the subgame perfect equilibrium (SPE) of the sequential move game,

(400, 400, 400) is the equal split, and (600, 500, 100) is the proposal that comes closest to equalizing payoffs for the proposer and responder while minimizing the payoff to the dummy player. Hence, this game combines the ultimatum and dictator games in the sense that the proposer and the one who votes on the offer play an ultimatum game, while the proposer plays a dictator game with the dummy player.

Note that the typical results from the ultimatum-dictator games (see Roth (1995)) are that: in both games people send positive amounts. In the ultimatum game the mode is close to the equal split and very little is sent above that. The amounts sent in the dictator game tend to be substantially lower than in the ultimatum game. Finally, in the ultimatum game, offers far below the equal split are rejected more often than almost equal offers.

Bellemare et al. (2008) find, amongst many results, that proposer's beliefs do not correspond to responder behavior. Also, there is substantial heterogeneity, both in behavior and expectations, which cannot be accounted for by observable characteristics.

In the three-player ultimatum game of Güth et al. (2007), the findings are that the most common offers are the equal split, followed by the 2-way almost equal split, and then the SPE offer. Acceptance rates increase with the offer to the person voting and decrease when the proposer's share is increasing or the dummy's share is decreasing.

When it comes to demographics, there are three results common to both studies which stand out, all having to do with age. First, older subjects react more to inequities. In the ultimatum game, the effect of age on proposers is small, and mainly shifts the most and second most popular offers from almost equal in favor of the proposer to slightly unequal in favor of the responder. When it comes to responders, both young and old display, on average, a plateau where the probability of acceptance goes down past the almost-equal split point. . The slope, in both directions away from the peak, is more pronounced among older subjects. In the three-player ultimatum game, the frequency of equal splits increases with age, and the frequency of two-way almost equal splits decrease with age. On the responder side, the (negative) impact of almost every form of inequity in the distribution becomes more important with age. Second, in both studies, older subjects reject more. It doesn't matter what it is, even offers as close to equal as possible, older subjects are more likely to reject than younger ones. Third, once you control for age, behavior is similar to that observed in the lab. For the ultimatum game, the preference

parameter estimates of the young and highly educated are similar to those of Fehr and Schmidt's (1999) calibrated distribution based on lab experiments. For the three player game, controlling for age and education or even simply controlling for age, results are not different from those based on lab experiments reported in Güth and van Damme (1998). In fact, both papers find that the effect of age is more important than that of education.

Carpenter, Connolly, and Myers (2008) employ a modified dictator game in which subjects divide $100 between themselves and a charity of their choice (with a 10% chance that their choice will be selected). Some subjects are from a pool of volunteers for experiments at Middlebury College, while the community members were drawn from a sample of 2000 randomly drawn addresses in the state of Vermont.[26] The main finding is in line with the other two studies previously mentioned, namely, older people take less for themselves as dictators, and being a student is not a robust determinant of behavior, while age is.[27]

Besides Güth et al. (2007), others have used newspaper or magazine experiments: Thaler (1997) in the *Financial Times* (United Kingdom), Bosch-Domènech and Nagel (1997) in *Expansión* (Spain) and in the *Financial* Times (United Kingdom), and Selten and Nagel (1997) in *Spektrum der Wissenschaft*, the German edition of *Scientific American*. All of those experiments investigated the Beauty Contest Game. Results from these experiments are discussed in Bosch-Domènech et al. (2002). However, unlike Güth et al. (2007), those studies did not collect information about demographics so the impact of these variables cannot be assessed. The key result from these experiments is the presence of spikes at 33.33, 22.22, and 0, with the comments from participants most often describing a logic of iterated best reply. The authors also report that results from experiments with the standard college student subject pool in the laboratory replicate these features except for the spike at 0 which does not exist for the college students.

Three more topics that fall in the individual decisions category have been studied in representative samples. First, Huck and Müller (2012) consider the standard Allais paradox choices over lotteries using the previously mentioned CentERdata. They also

[26] Random except with respect to gender as this sample was part of a larger study where males needed to be oversampled.
[27] Bekkers (2007) also reports (for a large Dutch sample) that donations to a charity in a dictator game increase with age, but that overall donations are very low.

conduct an experiment using the standard subject pool at the University of Tilburg. The experiment is a between subjects design with high hypothetical payoffs (millions of Euros), low hypothetical payoffs (up to 25 Euros), and low real payoffs. The results can be summarized as follows, in both representative and student samples, Allais type violations are much more common under high hypothetical payoffs than under low hypothetical payoffs (a difference of 28.2 and 30.6 percentage points for representative and students respectively, both are statistically significant). Moving from hypothetical small to real small payoffs produces only a small increase in violations (a difference of 4.4 and 3.1 percentage points for representative and students respectively, with only that for the representative sample being significant).[28] Hence, the comparative statics move in the same direction. One important difference, however, is that the number of violations is about 15 percentage points higher for the representative sample in all treatments. The demographics that affect behavior are: having a university degree (decrease violations), being employed or self-employed (decrease violations), and having higher income, savings, and assets all decrease violations.

Two other topics that have been investigated using representative samples are risk and intertemporal preferences.[29] Andersen et al. (2010) uses various modified versions of the Holt-Laury procedure to elicit risk preferences and modified versions of the Coller-Williams procedure to elicit intertemporal preferences.[30] They also vary the endowment given to subjects and the order in which various tasks are performed. These are done using students recruited at the University of Copenhagen and the Copenhagen Business School (100 subjects between the ages of 18 and 32) and using a demographically diverse sample of the Danish population (253 subjects between the ages of 19 and 75) that was

[28] Note that the representative sample is more than six times the size of the student sample and thus the test has much more power. In fact, when the representative sample is broken down into subgroups of demographics (by age groups, education level, etc.) the difference is no longer significant in 19 of the 29 cases considered.

[29] Beside the studies described here, I am aware of two other studies of risk preferences in a representative sample: Dohmen et al. (2011) and Harrison et al. (2007). The first uses a representative sample of the German population but their focus is on correlating a survey measure of risk to the standard experimental measure of risk. The second is based on the same sample as in Andersen et al. (2010) that I cover here in more detail.

[30] Both methods ask subjects a series of binary choices (Coller and Williams (1999) and Holt and Laury (2002)).

sampled to be representative.[31] The main results are the following: First, there are no statistically significant differences in the average degree of risk aversion between the two samples. The mean coefficient of relative risk aversion (CRRA) is estimated to be 0.63 in the representative sample and 0.79 in the student group. Second, there are no statistically significant differences in average discount rates across the two samples pooled across all horizons (1, 4, and 6 months). Subjects in the representative sample have an average discount rate of 25%, while students average 27.9%. Third, focusing on the representative sample, they identify that adults above 40 are less risk averse than those below, skilled workers are more risk averse, and students are more risk averse. It is somewhat surprising that these factors are significant, but that the average CRRA comes out to be almost the same in the student and representative samples. One potential explanation could be that, although the differences cited above have an impact on risk preferences, they explain but a small fraction of the variance in the CRRA coefficient. Fourth, again for the representative sample, none of the demographic variables (such as age, gender, being a parent, etc.) have a statistically significant impact on discount rates.

Another study relying on the CentERdata, as well as laboratory experiments by von Gaudecker et al. (2012) reports very different results. They use a modified Holt-Laury type procedure where subjects first choose among four lotteries (presented as pie charts) and then, if they exhibit at most a single switch point, they are presented with an additional screen of four lotteries that allow a more precise estimate of preferences. Payment is in three months, and for some options, the outcome is revealed immediately while for some others it is only revealed in three months. In addition, some choices can result in losses (from an endowment so that on net they cannot lose money). There was a high and low real payoff condition as well as a high hypothetical condition. In addition, sessions conducted in the laboratory are implemented in two ways: standard procedures with an experimenter present and one treatment that is closer in format to procedures employed with the CertER subjects. Among the few between sample results reported, one clear finding is that the standard subject pool makes much fewer mistakes (both with standard methods and with CertER type methods) than the representative sample. In the

[31] Although the sampling method was intended to generate a representative sample, the authors observe sample selection in the subjects who participated in the experiment.

standard subject pool, making at least one choice inconsistent with standard models happens 16.2 to 18.4 percent of the time. For the representative sample the frequency increases to 34.7 percent. The authors specify a structural model of choice that allows for a form of loss aversion and errors. They report that estimates of risk and loss aversion are smaller, on average, for the typical subject pool.

The results of Andersen et al. (2010) and von Gaudecker et al. (2012) are clearly at odds with these results. One possible explanation suggested by von Gaudecker et al. (2012) is that they have more data, and thus, more power to distinguish differences. Another possibility seems to be the difference in methods as Andersen et al. (2010) do not allow subjects to go back and forth in inconsistent ways and, in addition, offer subjects an option to express indifferences. Hence, the difference could be due to how the structural model fits subjects that make "mistakes". One could imagine, for instance, using the von Gaudecker et al. (2012) method, but explaining to subjects why it seems sensible to switch only once. Maybe then the CentERdata would look more like the standard subject pool. In other words, it is not clear that what is picked up by the structural model as a difference in risk preference is actually the expression of a preference.

I conclude this section with Belot et al. (2012), which does not use a representative sample, but compares students and non-students in a laboratory experiment.[32] They study the trust game, dictator game, a repeated VCM, a beauty-contest, a second-price auction, and simple choices between an amount for sure and various lotteries to determine risk aversion.[33] They report that students are closer to the equilibrium prediction with standard preferences in all games except for the auction, but that the differences are greater for games where other-regarding preferences have a potential role in that students are less other-regarding than the general population.[34] They exclude heterogeneity in risk preferences as a potential explanation for these differences based on the fact that both samples exhibit statistically similar behavior in that regard.

[32] The non-student subjects were recruited via e-mail to non-academic staff working at the university, by contacting local shops in Oxford, by placing advertisements in a local newspaper and in local pubs.
[33] Gächter et al. (2004) conducts a one-shot VCM experiment with students and non-students in rural and urban Russia. They find that non-students contribute more, but none of the multiple demographic variables they consider has a significant effect on contributions.
[34] See Cooper and Kagel (2014, Chapter XX of this volume) for comparisons of student samples with more representative samples with respect to other-regarding preferences.

They also exclude confusion on the basis of two facts. First, differences in behavior are greater in what they consider the simpler other-regarding preference related games (they view the dictator game as the simplest, then the trust game, and finally the VCM game). Second, when they control for instruction comprehension (they ask subjects to construct examples and to compute payoffs for those examples), the differences persist. I would point out, however, that when looking at the percentage of subjects who understood the instructions, that percentage is lower for non-students than for students in every single game.[35] Combining this with the fact that in the dictator, trust, and VCM games, mistakes are confounded with other-regarding preferences, suggests care in interpreting the results. Relatedly, Recalde et al. (2014) show, using the standard subject pool, evidence that suggests some donations in the VCM game may well be mistakes. When they modify the VCM game to have an interior solution, and that solution is for contributions to be low, they obtain the standard result that contributions are above what is predicted in equilibrium (with standard selfish preferences). However, when they modify the game such that the (interior) solution requires high contributions, then subjects on average under-contribute. Hence, by changing the environment they make confusion go in the direction opposite of what other-regarding preferences predict. This highlights the potential confound and is another indication of the possible increased confusion in the non-student subjects of Belot et al. (2012).

In the VCM game, for both students and non-students, Belot et al. (2012) find that repetitions decrease contributions and in both samples this happens at what appears to be a similar rate. Hence, although the difference in contributions persists, by the end (in round 10) the difference is no longer statistically significant, with the trend line indicating that any remaining differences will become negligible or completely disappear with more repetitions (since the choice space is bounded from below). The authors also point out that in terms of comparative statics; results are similar for students and non-students. In particular, for both samples, repetitions decrease contributions in the VCM and the frequency of equilibrium play is higher in what they view as the simpler games.

[35] I would also note that understanding the strategic considerations of a game is more than simply being capable of computing payoffs.

As a whole, there are a few patterns that seem to emerge from the use of demographically varied samples. One is that results are often not drastically different from those using the standard sample of student subjects, certainly when it comes to comparative statics, and to the extent that there are differences, these can often be traced to age. Another factor that sometimes has an impact is education. It is unclear, however, why some factors matter some time, for some behaviors, and not for others. There is no self-evident model of behavior or the role of specific demographic variables that one can see emerging from these studies taken as a whole. However, representative samples sometimes exhibit more "mistakes" (as compared to standard theoretical benchmarks). A couple of possible perspectives on this are: either that the standard subject pool gives a lower bound on the distance between behavior and standard models or that responses from representative samples are noisier. This could result from subjects having a harder time understanding instructions or because they are not incentivized enough.

Methodological Notes

Carpenter et al. (2008) point out that the dictator game is *unnatural* in the sense that it is a very artificial way of representing donations to charity. Although I agree with that statement, I would like to point out that originally, and in many other studies, the dictator game is not used as a way to understand charitable donations, but rather as a tool to decompose the strategic versus altruistic parts of positive offers in the ultimatum game (Forsythe et al. 1994). As such, it would seem undesirable to use it to gain insight into charitable giving.[36]

One potential issue with using surveys to administer experiments and with using non-students to take part in those experiments, is that there may be more confusion on the part of subjects. For instance, in Fehr et al. (2003), the amounts of money returned by the second movers in the trust game is only very weakly correlated to what the first movers sent; and in Güth et al. (2007), older people have a tendency to reject everything more than younger ones. Both of these observations could be the result of confusion.

Finally, newspaper experiments also have certain specific disadvantages. One is that there may be important selection effects in terms of who reads which paper. Also,

[36] Cooper and Kagel (2014, Chapter xx of this volume) detail the instabilities in the dictator game – while also pointing out its value for clarifying motives for behavior in the ultimatum game.

newspapers are sometimes unwilling to use neutral frames. In the case of the Güth et al. (2007) study, this resulted in presenting the problem as one of dividing money between brothers, which has its own specific set of moral judgments attached to it.

VII. Subjects With Relevant Task Experience

Experimental studies with professionals, subjects who have experience in the domain of interest, are clearly interesting. Professionals have experience at a task and have been selected in lieu of others to perform it. Take the case of auctions: it seems one could learn a lot from studying professional bidders and there is one such study that investigates the behavior of professionals from the construction industry in a common value auction. These subjects are interesting because they are the prototypical agent bidding in common value auctions. It is self-evident that people who bid for contracts for a living are auction professionals. However, are expert chess players the prototypical people playing the centipede game?[37] That seems less clear, as nobody actually plays the centipede game outside of the laboratory. This game was designed to illuminate certain aspects of strategic interactions: in particular backward induction. It is true that backward induction can be important in chess, and thus one would reason chess players must be good at it; but playing the centipede game is not something they are professionals at. Similarly, nurses are not professionals at playing the VCM game. They deliver health care, a public good, but hopefully they receive appropriate compensation, and thus they are not the ones bearing the cost of the public good. It may be that nurses are by nature more compassionate, but this is different from saying they are professionals in VCM type environments. That is why this section is about *subjects with relevant task experience*; not really professionals. I will use professionals as shorthand, but this is not to imply that these subjects necessarily have professional experience at the game being studied.

For the same reason expressed above, I believe our perspective on these experiments should be somewhat nuanced. If I learn that expert chess players all stop in the first move of the centipede game, that does not mean that professionals (businessman,

[37] In the centipede game, introduced by Rosenthal (1981), players take turns choosing either to stop the game or continue. At every node, the player who chooses would make less money if he continued and the other player immediately stopped; but he would make more money if the game continued past that point. The unique subgame perfect Nash equilibrium involves both players defecting at every node where it is their turn to choose.

traders, auction bidders, developers, bureaucrats, etc.) are rational. Students are not representative of the group of interest in some ways, but expert chess players are not representative either. Both groups are interesting, and our understanding of rationality, of the extent to which humans perform backward induction, as well as the role of common knowledge of rationality, depends on whether we find similar results or different results in these two samples. However, one does not invalidate the other, nor does a sample of such "professionals" provide more direct evidence about the rationality of people who engage in economic activity than the behavior of students.

The first study to compare professionals to the usual subject pool of undergraduate students in an economic experiment is, to the best of my knowledge, due to Fouraker et al. (1964). They compare bargaining game behavior of students with that of sales division employees from General Electric. However, the first paper of this sort to appear in an economics volume was published much later, in 1985, and it studied wool buyers in an oral double auction. This was followed by Dejong et al. (1988) comparing students to accounting or auditing partners and corporate financial officers in a principal-agent problem. The experiment by Fouraker et al. (1964) incentivized each group at different rates, although there is no mention of how these rates compared to the opportunity cost of their respective groups. The next two papers (Burns (1985) and Dejong et al. (1988)) had unusual incentives (either for both groups of subjects or at least for the group of professionals) and as such any difference in behavior could be the result of different incentive schemes. Beside these first three papers, the other papers that compare professionals to subjects in *standard* laboratory studies typically (with the exception of one) have one of two structures for compensation: students and professionals are paid the same or professionals are paid more than students.

Fréchette (2009) surveys in detail all studies that conduct treatments with undergraduate students and professionals using standard laboratory procedures with the goal of testing economic theories, which amounts to 13 papers.[38] Instead of covering all these papers in detail, I will summarize Burns (1985) as a representative of the early wave of studies (with unusual incentives). Then I will summarize a more recent paper by

[38] Fréchette (2009) is only one of the multiple chapters in Fréchette and Schotter (in press) discussing field and laboratory experiments. The interested reader is referred to Chapters 14 to 20 of that book.

Cooper, et al. (1999) that provides the most serious attempt at controlling incentives across groups. These two papers will illustrate one of the most interesting results to come out of this literature. The other papers will be very briefly described; the interested reader is referred to Fréchette (2009) for more details.[39] In addition, experiments with professionals that can be compared to experiments with students even if the two were not both conducted by the same authors or with the same procedures are covered here.[40]

Burns (1985) compares nine student subjects (second-year microeconomics undergraduates) to nine experienced wool buyers. They are both asked to bid as buyers on two units in a progressive oral auction with homogeneous commodities and fixed supply. The progressive oral auction with homogeneous commodities is close to the market the wool buyers operate in. As a way to stimulate trades, Burns introduced penalties for untraded units; she argues that the requirement to meet the full demand quota is very serious in the wool market. Fifteen auctions were conducted in total (five per week). Conditions were constant within week but demand changed across weeks.

The experiment was part of a course exercise for which an essay worth 10 percent of the students' final assessment must be written. The students did not know the subject of the essay, but they were advised that "only by striving to maximize their profits would they gain the understanding necessary to successfully complete the assignment." For the wool buyers, it was announced that "the 'best' trader would be revealed at the end of the session."

The data reveals that wool buyers bid up to their marginal values on the first lot, then marginal value plus penalty on the second. On the other hand, students behave similarly to wool buyer on day one of week one but then the demand curve flattens out in subsequent days (more contracts at or close to the market equilibrium prediction). Thus the students' behavior is closer to the competitive equilibrium prediction. As a result, students made much more money than wool buyers. When demand conditions changed, both students and wool buyers changed their behavior in the expected direction.

[39] In that piece I also explain aspects of the validity of an inference and elaborate on the reasons to study professionals.

[40] Some very interesting experimental studies of professionals that have no direct counterparts with students (either in topic or in their implementation) are not included here. These include, for instance: the study of social connections and the impact of managerial incentives in a fruit-picking farm (Bandiera (2009)), the possibility for gift-exchange in a tree-planting firm (Bellemare (2009)), gender discrimination by taxi drivers (Castillo (2013)), and many others.

Discussions and interviews with the wool buyers suggest that their behavior was driven by the behavior they have learned in the market they know. More specifically, wool is not a homogeneous good. Hence, these traders are not accustomed to noticing within-day price variations, as these can represent different quality wool. As a result, despite the fact that each auction featured a sharp decline in prices in the course of the session, seven of the nine professional buyers reported not noticing that pattern.

Cooper et al. (1999) compare Chinese students (10 sessions) to managers from the Chinese textile industry (12 sessions). They compared the behavior of these groups in a standard signaling game which is thought to represent a typical problem in centrally planned economies, namely the fact that production targets assigned by the central planner to specific firms increase with productivity (the ratchet effect), thus giving an incentive to firms to misrepresent their true productivity. Some sessions were conducted in generic terms, others employed meaningful context (referring to easy and tough contracts, and high-productivity firms, etc.). The game was repeated 36 times with roles reversed ("central planners" or "firms") after every six games.

The game has three pure-strategy sequential equilibria: pooling at output levels one, two, or three. The students had two payment schedules. In the standard pay: pooling at two corresponds to an expected payoff of 30 yuan ($3.75) in each game, which is equivalent to earnings in a typical U.S. experiment. In the high-pay cases, payments were multiplied by five, which represents 150 yuan in each game with pooling. As a point of comparison, the monthly wage for an associate professor was 1200 yuan. Managers had the same incentives as the students in the high-pay condition. Note that the vast majority of managers in the experiment earned less than 2000 yuan per month.

Overall, firm's choices start clustered around their type's full information output (they are not pooling, but rather acting myopically, not accounting for central planners responses to their choices). Central planners give easier contracts to outputs one to three than higher outputs. Experience increases the frequency of pooling by high productivity firms (strategic play) with play converging to output level two. However, a sizable frequency of nonstrategic play by high productivity firms remains even in the last 12 games. For students, increased pay promoted more strategic play by firms initially. However, by the end there were no differences. Increased pay had no impact on the

central planners' choices. Finally, there are no effects of context on students acting as firms. Mistakes by central planners are reduced for students but only in standard pay. There is an increased level of strategic play for managers in their role as firms in later cycles of play. There is a strong effect on managers as central planners, promoting higher target-rate differentials than in the generic sessions. To summarize, similar behavior is observed between students and managers. However, context helps managers come closer to the pooling equilibrium outcome.

The two studies above illustrate an important finding that emerges from the review in Fréchette (2009): to the extent that there are differences between professionals and students, these are often the result of aspects of the professional's work environment which are absent from the game being tested in the laboratory. The professionals either assume that features of their work environment are present in the lab when they are not or they rely on cues that can only be triggered by context. Consequently, these differences do not necessarily make the professionals behave more in line with the theory: sometimes they respond to elements relevant in their work but not in the particular setting being tested.

The papers reviewed in Fréchette (2009) are grouped under four broad headings: other-regarding preferences, market experiments, information signals, and a miscellaneous group. Other-regarding preferences includes four papers, the oldest being the bargaining experiment from Fouraker et al. (1964) using employees of the Industrial Sales Operation division of General Electric. For both professionals and students, when both sides are informed of the profits the other side makes, results are further from the equilibrium and closer to the equal split. Cadsby and Maynes (1998) compare nurses to students in a threshold public goods game.[41] The results in this case are very different. Subjects start with total contributions close to the threshold but as they gain experience their contributions average much lower than the threshold. Nurses on the other hand start high above the threshold and finish close to threshold. Fehr and List (2004) use CEOs from the coffee-mill sector in Costa Rica to play a trust game. The main results are that both CEOs and students send money when they are the first mover and ask for the second

[41] As opposed to a VCM game the public good is provided only if there is a certain level of contributions, this game has multiple equilibria, one with no contributions and many that have just enough contributions for the public good to be provided.

mover to send back less than the value of the amount sent (three times the original transfer). In both treatments, second movers send back money and in both cases it averages less than what the first mover asked for. CEOs differ from students in that they send more money when they are first movers (hence they are further away from the equilibrium prediction).[42] Carpenter and Seki (2010) study Japanese shrimp fishermen, some of who work in boats that share all expenses and revenue (poolers) and others who do not (non-poolers), in a VCM game. Both types of fisherman contribute more than students, with the contribution levels of the different types of fishermen not statistically different from one another.[43] Hence, as in Fehr and List (2004), professionals are further from the equilibrium than the students.

The other market study, besides Burns (1985), is DeJong et al. (1988). They study members of the professional Accounting Council of the Department of Accounting at the University of Iowa in a principal-agent problem. Principals offer a quality of service and price (sealed bid); agents see the offers and choose who to obtain their services from. The outcome of the service is determined randomly, but if the quality provided is too low, the loss needs to be covered by the principal. Prices, quality of services, and average expected profits are not statistically different when comparing professionals to students.

The category information signals include three studies testing models that rely on Bayesian updating. The first paper of this subgroup, Cooper et al. (1999), is covered above. The second, Potters and van Winden (2000), investigates the behavior of public-affairs and public-relations officers in a lobbying game: a signaling game with an informed sender and receiver. Many results are comparable for students and professionals, some in line with the theory and some not. Both groups react in the expected directions, given the strategic tensions. However, professionals in the role of senders are closer to equilibrium predictions in terms of how their messages vary with their information. Alevy et al. (2007) studies an information-cascade game with market professionals from the Chicago Board of Trade's floor. The results: a majority of choices in both samples are consistent with Bayesian updating; information cascades are realized at similar rates; and earnings are not different. However, professionals are slightly less

[42] There is also a treatment manipulation but it has only a small impact on behavior.
[43] The paper also explores a modification of the VCM game that allows subjects to express their emotions.

Bayesian than students. On the other hand, the behavior of students is sensitive to gains and losses, which is not the case for professionals.

The four remaining papers covered in Fréchette (2009) fall in the miscellaneous category. Dyer et al. (1989) compare the behavior of professionals from the construction industry (with at least 20 years of experience, many in bid preparation), to students' behavior in a first price sealed bid common value auction. Behavior of the two groups is similar in most key categories: average profits, percentage of times the winner is the one with the most optimistic signal, percentage of times the winning bid implies expected loses. In other words, both groups exhibit the winner's curse.

Palacios-Huerta and Volij (2008) investigate the behavior of professional soccer players in two zero-sum games (where the equilibria are in mixed strategies).[44] The argument for considering this group in this task is that penalty kicks require mixing. Both groups (soccer players and students) are fairly close to predicted behavior in terms of aggregate average choice frequencies (although the professionals are even closer than students). However, the way in which their behavior differs is in the independence of the choices across repetitions. Students, unlike soccer players, do not generate random sequences.

Abbink and Rockenback (2006) study bank employees from the departments of foreign exchange, security, futures, bonds, and money trade, in an option pricing experiment. Although students react more than professionals to a variable that should not affect their behavior (there are two risky states, the variable they react to is the probability of one state versus the other), their average behavior is closer to equilibrium. Also, with experience, students move closer to equilibrium while professionals move in the opposite direction. Finally, professionals do not arbitrage as much as students.

Cooper (2006) compares the behavior of undergraduates and executive MBAs in a weak-link game where *managers* can also set bonuses.[45] The bonus is costly to the managers but increases the value of efforts to the employee. Employees are always

[44] The soccer players are from the Spanish division one and division two. In addition to these and a sample of undergraduates with no soccer experience, they also have a sample of undergraduates who currently play in an official amateur senior regional league.

[45] In the weak-link, or minimum, game, employees pick costly effort and their profits increase as the minimum of the group increases such that any tuple of efforts that is identical for all employees is an equilibrium and these are Pareto ranked.

undergraduates and only managers vary. By the end, average minimum efforts, bonuses, and profits are similar for both professionals and students. However, professionals attain these levels quicker than students do.

On the basis of these 13 papers, I conclude that professionals do not seem qualitatively different from student in any systematic way. In most studies (nine out of the 13) they react to the forces at play in ways similar to students. In the cases where they differ, they are more often further away from the theory than closer. That is probably not because they are less sophisticated, but rather, as the first two studies summarized here highlight, because in their professional environment certain elements are present which are absent in the laboratory.[46]

There are other experiments with professionals that do not include student controls but share enough features with the typical procedures for them to inform us about robustness. In particular, the case of soccer players in zero-sum games (Palacios-Huerta and Volij (2008)) has been revisited in two papers. Wooders (2010) re-analyzes the data of Palacios-Huerta and Volij (2008). He finds that the behavior of soccer players follows non-stationary mixtures over the course of the experiment and that they tend to switch from underplaying to overplaying actions with respect to the minimax prediction. In those respects, and in terms of the distribution of action frequencies, he finds that students are actually closer to equilibrium than soccer players.[47]

Another related study is that of Levitt et al. (2010). They also study two zero-sum games (one which overlaps with Palacios-Huerta and Volij (2008)) on a sample of U.S. Major League Soccer players, a sample of professional poker players, and a sample of world-class bridge players.[48] Their result is that the behavior of all groups, students, poker players, bridge players, and soccer players, differs from the minimax predictions. This is true at the aggregate level, the individual level, and in terms of serial dependence.

Taking all three studies together, it is difficult to know what to think in terms of the ability of these professionals to behave optimally in these environments. For instance,

[46] Two more recent studies that include professionals and students are Roth and Voskort (2014) and Beck et al (2014). I will not describe either in detail as they apply to games that have not been widely studied. However, they both report qualitative similarities and quantitative differences between the two samples.
[47] The main difference in how the data is analyzed is that Wooders (2010) considers the first and second half of the data separately.
[48] They note that using mixed strategies has no role in bridge while it is an important part of poker.

it could be that Spanish soccer players are much better than American ones are.[49] There are some points worth making nonetheless. Although the three papers often mention how the frequency of play rejects the hypothesis that subjects mix in the correct proportions, in aggregate, these proportions are relatively close. Given the failures both within and across individuals to play mixed strategies, it is an open question as to what forces are at play that bring the aggregate proportions so close to their predicted values. However, it seems clear that the ability to randomize in a serially uncorrelated way is difficult for most subjects.[50]

Another group of professionals that has been the subject of study is chess players. Palacios-Huerta and Volij (2009) published a study of expert chess players recruited at international tournaments.[51] Their sample is composed of one time play of the centipede game at the tournament and of laboratory studies where subjects (both chess players and students) perform 10 plays of the centipede game. They find that in both settings, chess players' behavior, when playing other chess players, is extremely close to the theoretical prediction to take in the first node. In fact, when Grandmasters are the first players, the game always terminates at the first node. When playing repeatedly, chess players quickly converge to the equilibrium prediction. When chess players move first against students, the game terminates at the first node more often than when students play against students, but not as often as when they play against other chess players. Similarly, when students move first against chess players, the game terminates at the first node more often than when they play against other students. However, in a similar experiment, Levitt et al. (2011), who also recruited chess players at international tournaments, find very different results.[52] Their chess players are very similar to the standard subject pool of students. In fact, none of their Grandmasters ever stop at the first node. They also have subjects participate in a game that relies on backward induction but in which failure to follow backward induction has no impact on total payoffs (Race to 100) and find that 60 percent of the chess players behave in accordance with equilibrium in that game. Of the chess

[49] One difficulty in making this argument, however, is that in the Palacios-Huerta and Volij (2008) data, the behavior of the undergraduate students who were amateur players is, at least as close or closer to equilibrium than that of the professionals.
[50] Note that this is not an intrinsic feature of minimax, since a zero-sum game can be played only once.
[51] Chess players are ranked by Elo ratings. The difference in ratings between two players predicts the probability that each will win. Their sample is composed of players with ratings above 2000.
[52] Some of their players have Elo ratings below 2000, but most are above.

players consistent with backward induction in Race to 100, none stopped at the first node in the centipede game.

Two other studies use chess players, Bühren and Frank (2010) and de Sousa et al. (2014). In these, chess players play one or more beauty contest games. In both cases, results are far from the Nash equilibrium and close to the results (or even further from equilibrium) of experiments with students. In the case of de Sousa et al. (2014), this is true even in beauty contest games with only two players, where the solution is a dominant strategy.

Levitt et al. (2011) argue that professionals do not recognize the strategic similarities between the centipede game and chess, and this is why it is not surprising that they do not behave in line with equilibrium. To wit, when the game is more directly about backward induction (as in their Race to 100), their behavior is more often in line with equilibrium. This explanation seems difficult to reconcile with the behavior of chess players in the de Sousa et al. (2014) paper where, in a simple game, even when high Elo players know they are facing other high Elo players, and furthermore, even when the game has a dominant strategy, chess players behave similarly to students and relatively far from the equilibrium.

It is difficult to know what to take away from these varied results except, maybe, that chess players are professionals at playing chess more than anything else. There seems to be no easy way to reconcile the results of Palacios-Huerta and Volij (2009) and Levitt et al. (2011) except maybe that the discrepancy lies in the details of how the experiments were implemented. In some ways, the results of Bühren and Frank (2010) and de Sousa et al. (2014) seem more consistent with the behavior Levitt et al. (2011) report in the centipede game, but the same cannot be said of the behavior they observe in the Race to 100.

Looking at chess players and soccer players together, I find that explanations for the poor performance of professionals in the lab in certain games but not in others seem plausible in some, but not all, cases. If soccer players are professionals at playing mixed strategies with penalty kicks, but cannot employ mixed strategies in an experiment because the environment does not call on their understanding of randomizing, does it mean that if we simply frame the environment as one of a soccer game, then they would

get it right? This seems unlikely. Given that much of the failures in this particular case have to do with serial dependence, I would speculate that in the field, data sets do not provide as tight a test of players' ability to generate serially independent draws. Specifically, it would seem that long sequences of uninterrupted repeated plays in stationary environments between two players over a short period of time does not happen in the field. If anything, the studies with chess and soccer players raise as many questions about what allows professionals to perform well in the field as it does about the ways in which the behavior of professionals and students differ or not in the laboratory.

Multiple papers (by John List and coauthors) have explored the behavior of professional sportscards dealers. In List and Lucking-Reiley (2000) they compare the behavior of these dealers (they also study card collectors) in multi-unit auctions, either using a uniform-price auction or a Vickrey auction. The most directly relevant study with student subjects is Kagel and Levin (2001) who compare behavior in the uniform-price auction and in a dynamic version of the Vickrey auction.[53] As predicted by theory, the Vickrey auction yields more demand reduction (the difference in bids between the first and second unit) than the uniform-price auction for all types of subjects: dealers, collectors, and students. However, unlike Kagel and Levin (2001), List and Lucking-Reiley (2000) also observe some overbidding by professionals on the first unit in the uniform-price auction. Although these results are interesting, Kagel and Levin show that there are complicating factors to be considered (for a thorough analysis read Kagel and Levin (2014) Chapter XX of this volume).

List (2003 and 2004) studies the endowment effect, also with sportscard dealers and collectors, as well as with pin collectors. The first paper is done with sportscards, the second with other goods. The endowment effect, the tendency for individuals to value a good more once it is part of his endowment, has been documented in many studies with students.[54] Consistent with previous results, List observes a significant endowment effect overall. However, the extent of the endowment effect is negatively correlated to market

[53] The setting in Kagel and Levin (2001) is simpler for bidders as each human bidder interacts with computerized opponents who have single-unit demand.
[54] However, Plott and Zeiler (2005) show that changing instructions and procedures can decrease or even eliminate the endowment effect.

experience. In particular, for subjects with high trading intensity, 11 or more trades per month, there is absolutely no evidence of an endowment effect.[55]

The behavior of sportscard show participants is investigated again in Gneezy et al. (2006). The paper documents a phenomenon that violates most models of risky choices, which they term the *uncertainty effect*: the valuation of a risky prospect is below the value of the worst possible outcome. This is first established in a series of experiments using the standard subject pool. The experiments with the professionals are administered as a Vickrey auction for a single baseball card (one superior and one inferior) or lotteries over the two cards. The experiment with professionals finds the same robust phenomenon: almost all lotteries they experiment with are valued less than the worst card.

Thus, it seems that in some cases, such as the sportscard traders with the endowment effect, behavioral anomalies are mitigated by experience. However, taking the evidence of professionals as a whole, in many instances, what one would conclude from using students is qualitatively similar to what is observed with professionals.

Methodological Notes

Experiments with professionals can be insightful, but they also pose challenges.[56] There is a tendency to assume professionals should confirm standard models, that market forces should lead professionals to be unbiased optimizers. That is why some worry that student subjects are uninformative (as they do not have the relevant experience and have not been subject to market forces). Hence, when students and professionals behave the same way in the lab, and it is not in line with our standard models, some conclude that this is because transfer of knowledge across domains is difficult. But then, how do we make sense of the cases where professionals with market experience do behave closer to the standard model than students?[57] Why can behavior be transferred in those situations? I think this approach is not particularly useful and rests on problematic assumptions. For

[55] Although this is for all samples combined, dealers are the ones with the highest trading intensity. The papers do not report how many years of experience subjects who trade 11 or more times per month have, but from the summary statistics it would seem to be multiple years.
[56] See the multiple chapters discussing the field and lab experiments in Schotter and Fréchette (in press).
[57] Even if professionals do not transfer knowledge, how do we make sense of the cases where professionals are further away from equilibrium and make less money on average than students?

many of the professionals studied, it is unclear to what extent market-like forces correct biases. Think of nurses or chess players: what they do is complicated, and what features of their behavior are rewarded are not necessarily the ones that would bring their behavior in line with some simple optimizing and unbiased behavior. For example, Pope and Schweitzer (2011) establish that the behavior of professional golfers (including the very best) on the PGA tour exhibit loss aversion. If one accepts that biases (relative to expected utility theory) can exist even amongst top athletes, then observing violations of minimax by professional soccer players in the lab can be interpreted differently. Maybe the problem is not learning transfer, but rather that the laboratory offers a more stringent test of minimax because it is more precise and observations are better. In addition, in the case of many professionals, the fact that they are professionals does not mean they are necessarily highly sophisticated.

This is all to say that we might learn more by exploring the source of differences between professionals and students. As I discuss in Fréchette (2009), following the study of professional bidders in the construction industry where Dyer et al. (1989) observed the winner's curse, Dyer and Kagel (1996) interviewed the professionals to understand how it could be that they were successful businessman and yet lose money in their experiment.[58] An important finding is that the market they operate in is organized in ways to mitigate problems arising from the winner's curse. Hence, the problem is important enough to affect the industry in how it operates. If instead, these authors had assumed that businessmen could not fall prey to the winner's curse and that the issue must simply be an inability to transfer their knowledge; we would have missed an important discovery and the confirmation that the winner's curse is difficult to overcome.

VIII. Discussion

This discussion will be used to bring together results across subject pools on specific topics that have been studied in multiple samples. First it will cover topics in

[58] In the same spirit, Burns (1985) explores the reasons for the professionals' behavior in her experiment.

individual decisions and then move to games, many having some connection to other-regarding preferences.[59]

Individual Choice

Overall, the data suggests that GARP is alive and well, but it also indicates that it is learned. Harbaugh et al. (2001), who include a treatment with undergraduate students, report that in the standard subject pool, the vast majority of choices are in line with GARP.[60] Similarly, they also report that young children (ages seven and 11) display relatively low rates of violations, although the violations decrease with age. This result is confirmed in List and Millimet who also find that, using a sample of subjects between six and 17 years old, violations of GARP decrease with age. Adults in a token economy display choices for the most part consistent with GARP (Battalio et al. (1973)). Finally, the behavior of rats, pigeons, and monkeys is broadly consistent with GARP (Kagel et al. (1975), Battalio et al. (1981), and Chen et al. (2006)), the exception being pigeons who face multiple price changes, although they do move in the right direction.

Procedures and tasks vary so much between experiments considering risk preferences that it is difficult to draw general conclusions, in particular given the focus on point estimates (as opposed to comparative statics). Looking at rats (Battalio et al. (1985)), samples of adults of various ages (including the standard subject pool), as well as demographically varied subjects (Kovalchik et al. (2009), Kume and Suzuki (2012), Charness and Villeval (2009), and Harrison et al. (2007)), reveals a general tendency towards risk aversion, both across species and within humans; but no clear patterns relating age to risk preferences emerges.

Violations of EUT in the Allais paradox are observed in multiple samples: non-human animals (Kagel et al. (1990) and MacDonald et al. (1990)), in representative samples, and in the standard subject pool (Huck and Müller (2012)). However, the frequency of these violations is much lower in humans when the amounts are low, as compared to the original Allais example in the millions. Furthermore, subjects with a university education exhibit fewer violations.

[59] See Cooper and Kagel (2014, Chapter XX of this volume) for comparisons of student samples with more representative samples with respect to other-regarding preferences.

[60] The reader interested in a more complete review of the empirical evidence on GARP is referred to Varian (2006) and Andreoni et al. (2013).

I will loosely discuss in the same paragraph prospect theory, reference dependence, loss aversion, and the endowment effect. Although these are distinct concepts, prospect theory, which includes reference dependence and loss aversion, is often given as an explanation of the endowment effect. Evidence consistent with prospect theory and the endowment effect has been documented in multiple studies using undergraduate students (see Barberis (2013) for a review of some of this evidence). However, as we noted previously, Plott and Zeiler (2005) show that evidence on the endowment effect is sensitive to procedures. Notwithstanding this, overall there seem to be multiple instances of behavior consistent with prospect theory and the endowment effect. In line with this, Chen et al. (2006) find evidence consistent with both reference dependence and loss aversion in monkeys. Harbaugh et al. find the endowment effect in Kindergarteners, third graders, fifth graders, and undergraduate students. Furthermore, it is exhibited to the same extent at all ages. On the other hand, using Plott and Zeiler's (2005) procedures, Kovalchik et al. (2006) find no evidence of the endowment effect in college students nor in adults between 70 and 95 years old. List (2003 and 2004) observes the endowment effect, but it disappears in subjects with intense market experience.[61]

Overall, it seems that there is a tendency, shared across species, to exhibit the endowment effect; but that procedures and market experience can make it go away. However, to put this result in perspective, it is interesting to mention the study of Englemann and Hollard (2010). They note that the endowment effect could be the result of subjects being uncertain about market procedures, a phenomenon separate from uncertainty over the value of an object. Their experiment introduces a treatment where subjects are forced to trade in three consecutive rounds (otherwise they lose the value of the item they are endowed with) in order for them to learn about market procedures. Subjects who first go through these three rounds do not display the endowment effect.

From his experiments, List concludes, "This result is consistent with the notion that via previous market interaction and arbitrage opportunities, agents have learned to treat goods leaving their endowment as an opportunity cost rather than a loss." (List 2004, p. 624) First, note that, although this might be the case, market experience seems to

[61] Although not reviewed here, note also that List and Haigh (2010) find that the behavior of futures and options pit traders from the Chicago Board of Trade is closer to myopic loss aversion than the behavior of students.

be a rather inefficient teacher as it takes only three trials in the lab to get as much experience as professionals with years of intense trading. Second, the results of Plott and Zeiler (2005) and of Englemann and Hollard (2010) taken together point instead towards individuals learning about market procedures. This is not to say that the endowment effect is not an important phenomenon, but rather that it might be the expression of something different from what it is typically assumed to be.

With respect to time preference, two findings are robust across subject pools. First, Andersen et al. (2010) finds that discount rates are not statistically different between students and a representative sample of the Danish population. Second, Kagel and Green (1987) and Bettinger and Slonim (2007) report behavior consistent with hyperbolic discounting in pigeons and children, respectively. Many studies have observed hyperbolic discounting in student samples; however, some have argued that hyperbolic discounting for laboratory choices over money make little sense given the fungibility of money. In addition, what might look like present-bias could simply be uncertainty over future payments.[62] We note that neither of these concerns applies to animals, and the fact that adults can move money around probably does not apply to young children. Hence, overall, hyperbolic discounting seems to be present from a young age, and in other animals beside humans.

Games

In the ultimatum and dictator game, results are similar across samples, namely children (Harbaugh et al. (2003)), representative samples (Bellemare et al. (2008)), and professionals (Fouraker et al. (1964)). Just like standard subjects (Roth (1995)), they all make offers closer to the equal split than predicted by the subgame-perfect equilibrium in the ultimatum game. Also, as offers decrease, the probability of rejection increases. Furthermore, offers are higher in the ultimatum game than in the dictator game. Harbaugh et al. (2003) suggest that this behavior is, at least in part, learned as both offers and the rejection of low offers increase with age.

[62] See, for instance, Augenblick et al. (2013) for a discussion of these issues and an experiment that addresses them where they show hyperbolic discounting over non-monetary rewards.

In the VCM game, the standard results are: above-minimum contributions that react to the MPCR and are decreasing with repetitions (Ledyard (1995)). A similar pattern is exhibited by older children (above 11.5 years in Harbaugh and Krause (2000) and nine to 16 years old in Peters et al. (2004)). However, in children under 11.5 years old, Harbaugh and Krause (2000) do not find decreasing contributions when the task is repeated. The above-minimum contributions that decreases with repetitions is also observed in a sample of non-student subjects (Belot et al. (2012)) and amongst different types of professionals: the junior and senior workers in manufacturing in Charness and Villeval (2009) as well as in the fishermen who do not pool resources in Carpenter and Seki (2010). For fishermen who pool revenue and expenses, contributions are above the minimum, but show no signs of decrease with repetitions (the source of the difference in this case could very well be selection). Overall, behavior seems consistent in most samples in the VCM game, however some aspects of it might be learned as indicated by the fact that contributions do not decrease for young children.

In the trust game, though there are quantitative variations, many of the qualitative results carry over between populations. First, whether it is children (Harbaugh et al. (2002)), older subjects (Holm et al. (2005)), representative samples (Fehr et al. (2003), Bellemare and Kroger (2007), and Fehr et al. (2013)), or amongst CEOs (Fehr and List (2004)), on average, a positive amount is sent by the first mover, just like results with undergraduates. Similarly, in all groups, the amount sent back by the second mover is a function of what the first mover gave. Although it is difficult to establish clearly in all of these papers, from the summary statistics and figures, it seems as if sending more as a first mover increases the return but not enough, on average, to compensate for the loss. Taken together the evidence seems to point to a concave relation between age and the amount sent by the first mover. The impact of age on second movers is not as clear, some find that it matters, but do not find a consistent relation; others report no differences with age.

Some aspects of the beauty contest game seem to generalize across samples. In particular, be it older people (Kovalchik et al. (2005)), diverse subjects via newspapers (Selten and Nagel (1998), Thaler (1997), and Bosch-Domènech et al. (2002)), or chess players (Bühren and Frank (2010) and de Sousa et al. (2014)), in no case is the winning

number ever the equilibrium of zero. Furthermore, many authors report spikes at 22.22 and 33.33 (some do not mention if there are spikes or not). All of these patterns are consistent with what is observed in the standard subject pool (Bosch-Domènech et al. (2002) and Kovalchik et al. (2005)). However, the newspaper experiments also reveal a spike at zero and a lower winning number: between 12 and 17 depending on the study as opposed to about 24 in the typical experiment. For chess players, Bühren and Frank (2010) report a winning number of 21.43 with the guesses of the grandmasters slightly above the average for all players. This suggests that time to think (as in the newspaper experiment) might affect responses more than the intellectual sophistication of the subjects.

IX. Conclusion

In most of the cases for which a task or game has been studied in multiple samples, the results are surprisingly consistent. Qualitative results and comparative statics often carry over from professionals to other species, to undergraduate students. Of course there are some exceptions, but they seem to be just that, exceptions to the general rule. In some cases, certain behavior seems to evolve in childhood. In other cases, behaviors are affected by market exposure. However, the effect of market exposure can be replicated fairly quickly in the laboratory, suggesting it is not about market selection, but simply about experimentation.

When it comes to point estimates, moving across samples the results are a lot more diverse. However, even with respect to point estimates, it is surprising to find that in some cases there are far fewer (or less obvious) differences than one would have expected -- for instance, the similar risk taking behavior for young and old adults.

There is no doubt much to be learned from exploring samples besides undergraduate students. In particular, such studies are needed to understand to what extent our typical subject pool puts limits, if any, on the type of questions we can explore with some level of confidence and on the factors that affect the robustness of the results. This being said, looking at these studies together gives more confidence than worry about what can be learned from the typical subject pool. In particular, cases where treatment effects (or comparative statics) are different when considering the standard subjects as

opposed to other subjects are extremely rare. This also underscores the desirability of focusing on comparative statics. In particular, some of these studies remind us that using the laboratory to estimate preference parameters in designs where mistakes and non-standard preferences move in the same direction (away from predictions for perfectly optimizing and selfish agents) can be problematic. This can be a serious problem, especially when interpreting results across subject pools.

In some ways, these studies highlight many of the advantages from using student subjects: They make replication easier, are less costly, and are easily accessible. Students, by their training, find it easier to understand abstract written instructions. Finally, students, unlike professionals, are less likely to import irrelevant experiences and heuristics into the study, factors that may matter in the setting they typically operate in, but not in the environment under investigation.

References

Abbink, Klaus, and Bettina Rockenbach. 2006. "Option Pricing by Students and Professional Traders: A Behavioural Investigation." *Managerial and Decision Economics*. 27 497-510.

Alevy, Johnatan E., Michael S. Haigh and John A. List. 2007. "Information Cascades: Evidence from a Field Experiment with Financial Market Professionals." *Journal of Finance*, Vol. 62(1), 151-180.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, E. Elisabet Rutstrom. 2010. "Preference Heterogeneity in Experiments: Comparing the Field and Laboratory." *Journal of Economic Behavior & Organization*, 73 (2010) 209–224.

Andreoni, James, Benjamin J. Gillen and William T. Harbaugh. 2013. "The Power of Revealed Preference Tests: Ex-Post Evaluation of Experimental Design". *Working Paper*.

Augenblick, Ned, Muriel Niederle and Charles Sprenger. 2013. "Working Over Time: Dynamic Inconsistency in Real Effort Tasks," *Working Paper*.

Ayllon T. and N. H. Azrin. 1965. "The Measurement and Reinforcement of Behavior of Psychotics." *Journal Experimental Analysis of Behavior*, 8(6): 357-383.

Ball, Sheryl B. and P. A. Cech. 1996. "Subject Pool Choice and Treatment Effects in Economic Laboratory Research," in *Research in Experimental Economics*, vol. 6 (R.M. Isaac, ed.), 232-292.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2009. "Social Connections and Incentives in the Workplace: Evidence from Personnel Data," *Econometrica*, 77(4):1047-1094.

Barberis, C. Nicholas. 2013. "Thirty Years of Prospect Theory in Economics: A Review and Assessment," *Journal of Economic Perspectives*, 27, 173-196.

Basmann, Robert L., Raymond C. Battalio, and John H. Kagel. 1976. "An Experimental Test of a Simple Theory of Aggregate Per-capita Demand Functions." *Schweizerische Zeitschrift für Volkswirtschaft und Statistik / Revue suisse d'Economie politique et de Statistique*, 112(2) : 153-173.

Bateson, Melissa, Susan D. Healy, and T. Andrew Hurly. 2003. "Context-Dependent Foraging Decisions in Rufous Hummingbirds." *Proceedings of The Royal Society*, 270: 1271-1276.

Battalio, Raymond C, Gerald P. Dwyer, Jr, and John H. Kagel. 1987. "Tests of Competing Theories of Consumer Choice and the Representative Consumer Hypothesis." *The Economic Journal*, 97(388): 842-856.

Battalio, Raymond C., Edwin B. Fisher Jr., John H. Kagel, Robert L. Basmann, Robin C. Winkler, and Leonard Krasner. 1974. "An Experimental Investigation of Consumer Behavior in a Controlled Environment." *The Journal of Consumer Research*, 1(2): 52-60.

Battalio, Raymond C., Leonard Green, and John H. Kagel. 1981. "Income-Leisure Tradeoffs of Animal Workers." *American Economic Review*, 71(4): 621-32.

Battalio, Raymond C. and John H. Kagel. 1985. "Consumption-Leisure Tradeoffs of Animal Workers: Effects of Increasing and Decreasing Marginal Wage Rates in a Closed Economy Experiment," in *Research in Experimental Economics* (Vernon L. Smith, ed.), Vol. 3. Greenwich, CT: JAI Press: 1-30.

Battalio, Raymond C., John H. Kagel, and Leonard Green. 1979. "Labor Supply of Animal Workers: Towards an Experimental Analysis," in *Research in Experimental Economics* (Vernon L. Smith, ed.), Greenwich, CT: JAI Press.

Battalio, Raymond C., John H. Kagel, and Carl A. Kogut. 1991. "Experimental Confirmation of the Existence of a Giffen Good." *American Economic Review*, 81(4): 961-70.

Battalio, Raymond C., John H. Kagel, Howard Rachlin, and Leonard Green. 1981. **"**Commodity-Choice Behavior with Pigeons as Subjects." *The Journal of Political Economy*, 89(1): 67-91.

Battalio, Raymond C., John H. Kagel, and Don N. MacDonald. 1985. "Animals' Choices over Uncertain Outcomes: Some Initial Experimental Results." *American Economic Review*, 75(4): 597-613.

Battalio, Raymond C., John H. Kagel, and Morgan O. Reynolds. 1977. "Income Distributions in Two Experimental Economies." *The Journal of Political Economy*, 85(6): 1259-1271.

Battalio, Raymond C., John H. Kagel, and Morgan O. Reynolds. 1978. "A Note on the Distribution of Earnings and Output Per Hour in an Experimental Economy." *The Economic Journal*, 88(352): 822-829.

Battalio, Raymond C., John H. Kagel, Robin C. Winkler, Edwin B. Fisher JR., Robert L. Basmann, and Leonard Krasner. 1973. "A Test Of Consumer Demand

Theory Using Observations Of Individual Consumer Purchases." *Western Economic Journal*, 11(4): 411-428.

Beck, Adrian, Rudolf Kerschbamer, Jianying Qiu, and Matthias Sutter. 2014. "Car Mechanics in the Lab - Investigating the Behavior of Real Experts on Experimental Markets for Credence Goods." *Working Paper.*

Becker, Gary S. 1962. "Irrational Behavior and Economic Theory." *The Journal of Political Economy*, 70(1): 1-13.

Becker, Gordon M., Morris H. DeGroot, and Jacob Marschak. 1964. "Measuring Utility by a Single-Response Sequential Method." *Behavioral Science*, 9(3): 226-232.

Bekkers, Rene. 2007. "Measuring Altruistic Behavior in Surveys: The All-or-Nothing Dictator Game." *Survey Research Methods*, 1(3), 139-144.

Bellemare, Charles and Sabine Kroger. 2007. "On Representative Social Capital," *European Economic Review*, 51(1), 183-202.

Bellemare, Charles, Sabine Kröger, and Arthur van Soest. 2008. "Measuring Inequity Aversion in a Heterogeneous Population Using Experimental Decisions and Subjective Probabilities." *Econometrica*, 76(4): 815-839.

Bellemare, Charles, and Bruce Shearer. 2009. "Gift Giving and Worker Productivity: Evidence from a Firm Level Experiment," *Games and Economic Behavior*, 67(1): 233-244.

Belot, Michele, Raymond M. Duch, and Luis M. Miller. 2012. "Who Should Be Called to the Lab: A Comprehensive Comparison of Students and non-Students in Classic Experimental Games," *Working Paper.*

Besedeš, Tibor, Cary Deck, Sudipta Sarangi, and Mikhael Shor. 2012. "Age Effects and Heuristics in Decision Making." *The Review of Economics and Statistics*, MIT Press, vol. 94(2), pages 580-595, May.

Bettinger, Eric and Robert Slonim. 2006. "Using Experimental Economics to Measure the Effects of a Natural Educational Experiment on Altruism." *Journal of Public Economics*, 90:1625-1648.

Bettinger, Eric and Robert Slonim. 2007. "Patience Among Children." *Journal of Public Economics*, 91: 343-363.

Bolton, Gary E., and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review*, 90(1): 166-193.

Bosch-Domènech, Antoni and Rosemarie Nagel. 1997. "El Juego de Adivinar el Numero X: Una Explicacion y la Proclamacion del Vencedor." *Expansion*, June 16, pp. 42-43.

Bosch-Domènech, Antoni, José G. Montalvo, Rosemarie Nagel, and Albert Satorra. 2002. "One, Two, (Three), Infinity, ...: Newspaper and Lab Beauty-Contest Experiments." *American Economic Review*, 92(5): 1687-1701.

Bühren, Christoph, and Björn Frank. 2010. "Chess players' performance beyond 64 squares: A case study on the limitations of cognitive abilities transfer." *Working Paper*.

Burns, Penny. 1985. "Experience and Decision Making: A Comparison of Students and Businessmen in a Simulated Progressive Auction," in *Research in Experimental Economics*, Vol. 3, Greenwich, CT: JAI Press: 139-153.

Cadsby, C. Bram and Elizabeth Maynes. 1998. "Choosing Between a Socially Efficient and Free-Riding Equilibrium: Nurses Versus Economics and Business Students." *Journal of Economic Behavior & Organization*, 37(2), 183-192.

Carpenter, Jeffrey, Cristina Connolly, and Caitlin Myers. 2008. "Altruistic Behavior in a Representative Dictator Experiment." *Experimental Economics*, 11(3): 282-298.

Carpenter, Jeffrey, and Erika Seki. 2010. "Do Social Preferences Increase Productivity? Field Experimental Evidence from Fishermen in Toyama Bay." *Economic Inquiry*, Volume 49, Issue 2, pages 612–630.

Castillo, Marco, Ragan Petrie, Maximo Torero, Lise Vesterlund. 2013. "Gender Differences in Bargaining Outcomes: A Field Experiment on Discrimination," *Journal of Public Economics*, Forthcoming.

Charness, Gary, Guillaume R. Fréchette, and John H. Kagel. 2004. "How Robust is Laboratory Gift Exchange?" *Experimental Economics*, 7(2): 189-205.

Charness, Gary and Marie-Claire Villeval. 2009. "Cooperation and Competition in Intergenerational Experiments in the Field and the Laboratory," *American Economic Review*, 99(3): 956-78

Chen, M. Keith, Venkat Lakshminarayanan, and Laurie R. Santos. 2006. "How Basic Are Behavioral Biases? Evidence from Capuchin Monkey Trading Behavior." *Journal of Political Economy*, 114(3): 517-537.

Coller, Maribeth and Melonie B. Williams. 1999. "Eliciting Individual Discount Rates," *Experimental Economics*, Volume 2, Issue 2, pp 107-127.

Cooper, David J. 2006. "Are Experienced Managers Experts at Overcoming Coordination Failure?" *The B.E. Journal of Economic Analysis & Policy*, Volume 5, Issue 2.

Cooper, David J., **John H. Kagel and Wei Lo and Qing Liang Gu.** 1999. "Gaming Against Managers in Incentive Systems: Experimental Results with Chinese Students and Chinese Managers." *American Economic Review*, 89: 781-804.

Cox, James C. 1997. "On Testing the Utility Hypothesis." *The Economic Journal*, 107(443): 1054-1078.

de Sousa, Jose, Guillaume Hollard and Antoine Terracol. 2013. "Non-strategic players are the rule rather than the exception." *Working Paper*.

Dejong, Douglas V., **Robert Forsythe, and Wilfred C. Uecker.** 1988. "A Note On The Use of Businessmen As Subjects In Sealed Offer Markets." *Journal of Economic Behavior and Organization*, 9: 87-100.

Dohmen, Thomas J, Armin Falk, David Huffman, Jürgen Schupp, Uwe Sunde, and Gert Georg Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants and Behavioral Consequences." *Journal of the European Economic Association*, 9(3): 522-550.

Dyer, Douglas, John H. Kagel, and Dan Levin. 1989. "A Comparison of Naïve and Experienced Bidder in Common Value Offer Auctions: A Laboratory Analysis." *The Economic Journal*, 99. 108-115.

Englemann, Dirk, and Guillaume Hollard. 2010. "Reconsidering the Effect of Market Experience on the "Endowment Effect"." *Econometrica*, 78(6), 2005–2019.

Dyer, Douglas and John H. Kagel. 1996. "Bidding in Common Value Auctions: How the Commercial Construction Industry Corrects for the Winner's Curse". *Management Science*, 42(10), 1463-1475.

Falk, Armin, Stephan Meier, and Christian Zehnder. 2013. "Do Lab Experiments Misrepresent Social Preferences? The Case of Self-Selected Student Samples." *Journal of the European Economic Association*, Volume 11, Issue 4, pages 839–852.

Fehr, Ernst, Urs Fischbacher, Jürgen Schupp, Bernhard von Rosenbladt, and Gert Georg Wagner. 2003. "A Nationwide Laboratory Examining Trust and Trustworthiness by Integrating Behavioural Experiments into Representative Surveys." IEW *Working Paper* No. 141.

Fehr, Ernst, and John A. List. 2004. "The Hidden Costs and Returns of Incentives – Trust and Trustworthiness among CEOs." *Journal of the European Economic Association*, 2(5), 743-771.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*. 114(3): 817-868.

Forsythe, Robert, Joel L. Horowitz, N. E. Savin, and Martin Sefton. 1994. "Fairness in Simple Bargaining Experiments." *Games and Economic Behavior*, 6:347-369.

Fouraker, Lawrence E., **Sidney Siegel, and D. L. Harnett.** 1962. "An Experimental Disposition of Alternative Bilateral Monopoly Models Under Conditions of Price Leadership." *Operations Research*, 10(1): 41-50.

Fréchette, Guillaume R. 2009. "Laboratory Experiments: Professionals versus Students," in *The Methods of Modern Experimental Economics*, Guillaume Fréchette and Andrew Schotter (editors), Oxford University Press, forthcoming.

Fréchette, Guillaume R., and Andrew Schotter (Editors). In Press. *The Methods of Modern Experimental Economics*, Oxford University Press.

Frederick, Shane, George Loewenstein and Ted O' Donoghue. 2002. "Time Discounting and Time Preference: a Critical Review." *Journal of Economic Literature*, 40, 350-401.

Gächter Simon, Benedikt Herrmann, Christian Thöni. 2004. "Trust, Voluntary Cooperation, and Socio-Economic Background: Survey and Experimental Evidence." *Journal of Economic Behavior and Organization*, 55, 505–531.

Gneezy, Uri, John A. List and George Wu. 2006. "The Uncertainty Effect: When a Risky Prospect is Valued Less than its Worst Possible Outcome." *Quarterly Journal of Economics*, 121 (4): 1283-1309.

Güth, Werner, Carsten Schmidt, and Matthias Sutter. 2007. "Bargaining Outside the Lab - a Newspaper Experiment of a Three-Person Ultimatum Game." *Economic Journal*, 117(518): 449-469.

Harbaugh, William T. and Kate Krause. 2000. "Children's Altruism in Public Good and Dictator Experiments." *Economic Inquiry*, 38(10): 95-109.

Harbaugh, William T., Kate Krause, and Timothy R. Berry. 2001. "GARP for Kids: On the Development of Rational Choice Behavior." *American Economic Review*, 91(5): 1539-1545.

Harbaugh, William T., Kate Krause, and Steven G. Liday Jr. 2003. "Trust Bargaining by Children." *Working Paper*.

Harbaugh, William T., Kate Krause, Steven G. Liday Jr., and Lise Vesterlund. 2002. "Trust in Children." *Working Paper.*

Harbaugh, William T., Kate Krause, and Lise Vesterlund. 2001. "Are Adults Better Behaved Than Children? Age, Experience, and the Endowment Effect." *Economics Letters*, 70(2): 175-181.

Harbaugh, William T., Kate Krause, and Lise Vesterlund. 2002. "Risk Attitudes of Children and Adults: Choices Over Small and Large Probability Gains and Losses." *Experimental Economics*, 5(1): 53-84.

Harbaugh, William T., Kate Krause, and Lise Vesterlund. 2007. "Learning to Bargain." *Journal of Economic Psychology*, Elsevier, 28(1), 127-142, January.

Harrison, Glenn W., Morten I. Lau, and E. Elisabet Rutström. 2007. "Estimating Risk Attitudes in Denmark: A Field Experiment." *Scandinavian Journal of Economics*, 109(2): 341-368.

Herrnstein, Richard J. 1961. "Relative and Absolute Strength of Responses as a Function of Frequency of Reinforcement." *Journal of the Experimental Analysis of Behaviour*, 4, 267–72.

Holm, Håkan and Paul Nystedt. 2005. "Intra-Generational Trust—a semi-Experimental Study of Trust Among Different Generations." *Journal of Economic Behavior & Organization*, 58: 403–419.

Holt, Charles A., and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review*, 92(5): 1644-1655.

Huck, Steffen, and Wieland Müller. 2012. "Allais for all: Revisiting the paradox in a large representative sample." *Journal of Risk and Uncertainty*, 44 (3) 261 - 293.

Kagel, John H. 1972. "Token Economies and Experimental Economics." *The Journal of Political Economy*, 80(4): 779-785.

Kagel, John H., **Raymond C. Battalio, and Leonard Green.** 1995. *Economic Choice Theory*. Cambridge University Press, 230 pages.

Kagel, John H., **Raymond C. Battalio, and C. G. Miles.** 1980. "Marihuana And Work Performance: Results From An Experiment." *The Journal of Human Resources*, 15(3): 373-395.

Kagel, John H., **Raymond C. Battalio, and James M. Walker**. 1979. "Volunteer Artifacts in Experiments in Economics: Specification of the Problem and Some Initial Data from a Small-Scale Field Experiment." *Research in Experimental Economics*, (1) 169-197.

Kagel, John H., Raymond C. Battalio, Howard Rachlin, and Leonard Green. 1981. "Demand Curves for Animal Consumers." *The Quarterly Journal of Economics*, 96(1): 1-16.

Kagel, John H., Raymond C. Battalio, Howard Rachlin, Leonard Green, Robert Basmann, and W. R. Klemm. 1975. "Experimental Studies of Consumer Demand Behavior Using Laboratory Animals." *Economic Inquiry*, 13(1): 22-38.

Kagel, John H., Raymond C. Battalio,, Robin C. Winkler, and Edwin B. Fisher, Jr. 1977. "Job Choice and Total Labor Supply: An Experimental Analysis." *Southern Economic Journal*, 44(1): 13-24.

Kagel, John H. and Leonard Green. 1987. "Intertemporal Choice Behavior: Evaluation of Economic and Psychological Models," in *Advances in Behavioral Economics* (John H. Kagel and Leonard Green Eds.), Volume 1, Ablex Publishing.

Kagel, John H., Don N. MacDonald, and Raymond C. Battalio. 1990. "Tests of "Fanning Out" of Indifference Curves: Results from Animal and Human Experiments." *American Economic Review*, 80(4): 912-921.

Kagel, John H., and Dan Levin. 2001. "Behavior in Multi-Unit Demand Auctions: Experiments with Uniform Price and Dynamic Vickrey Auctions." *Econometrica.* 69(2) 413-454.

Kume, Koichi and Ayako Suzuki. 2010. "Ageing, Probability Weighting, and Reference Point Adaptation: Experimental Evidence." *Working Paper*.

Kovalchik, Stephanie, Colin F. Camerer, David M. Grether, Charles R. Plott, and John M. Allman. 2005. "Aging and decision making: a comparison between neurologically healthy elderly and young individuals." *Journal of Economic Behavior & Organization*, 58: 79–94.

Ledyard, John O. 1995. "Public Goods: A Survey of Experimental Research." In *Handbook of Experimental Economics*, edited by John H. Kagel and Alvin E. Roth, Chapter 2.

Levitt, Steven D. and John A. List, David H. Reiley. 2010. "What Happens in the Field Stays in the Field: Exploring Whether Professionals Play Minimax in Laboratory Experiments." *Econometrica*, 78 (4).

Levitt, Steven D., John A. List, and Sally E. Sadoff. 2011. "Checkmate: Exploring Backward Induction among Chess Players." *American Economic Review*, 101(2): 975-90.

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

List, John A. 2004. "Neoclassical Theory versus Prospect Theory: Evidence from the Marketplace." *Econometrica*, 72(2), 615-625.

List, John A., and David Lucking-Reiley. 2000. "Demand Reduction in Multiunit Auctions: Evidence from a Sportscard Field Experiment." *American Economic Review*, 90(4): 961-972.

List, John A. and Daniel L. Millimet. 2008. "The Market: Catalyst for Rationality and Filter of Irrationality," *The B.E. Journal of Economic Analysis & Policy*, 8(1): 1-55.

List, John A. and Micheal S. Haigh. 2010, "Investment under Uncertainty: Testing the Options Model with Professional Traders." *Review of Economics and Statistics,* 92(4): 974–984.

MacDonald, Don N., John H. Kagel, Raymond C. Battalio. 1991. "Animals' Choices Over Uncertain Outcomes: Further Experimental Results." *The Economic Journal*, 101(408): 1067-1084.

Machina, Mark J. 1987. "Choice Under Uncertainty: Problems Solved and Unsolved," *Journal of Economic Perspectives*, 1, 121-54.

Madies, Thierry, Marie Claire Villeval, and Malgorzata Wasmer. 2013. "Intergenerational Attitudes Towards Strategic Uncertainty and Competition: A Field Experiment in a Swiss Bank." *European Economic Review*, 61, 53-168.

Nagel, Rosemarie and Reinhard Selten. 1997. "1000 DM zu gewinnen," in *Spektrum der Wissenschaft*, November.

Palacios-Huerta, Ignacio and Oscar Volij. 2008. "Experientia Docet: Professionals Play Minmax in Laboratory Experiments." *Econometrica*, 76(1), 71-115.

Palacios-Huerta, Ignacio, and Oscar Volij. 2009. "Field Centipedes." *American Economic Review*, 99(4): 1619-35.

Peters, H. Elizabeth, A. Sinan Ünür, Jeremy Clark, and William D. Schulze. 2004. "Free-Riding and the Provision of Public Goods In the Family: A Laboratory Experiment." *International Economic Review*, 45(1): 283-299.

Plott, Charles R. and Kathryn Zeiler. 2005. "The Willingness to Pay–Willingness to Accept Gap, the "Endowment Effect," Subject Misconceptions, and Experimental Procedures for Eliciting Valuations." *The American Economic Review*, 95(3),: 530-545.

Pope, Devin G., and Maurice E. Schweitzer. 2011. "Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes." *American Economic Review*, 101(1): 129-57.

Potters, Jan, and Frans van Winden. 2000. "Professionals and Students in a Lobbying Experiment Professional Rules of Conduct and Subject Surrogacy." *Journal of Economic Behavior & Organization*, 43, 499–522.

Rachlin, Howard, John H. Kagel, and Raymond. C. Battalio. 1980. "Substitutability in Time Allocation," *Psychological Review*, 87.

Rachlin, Howard, Raymond. C. Battalio, John H. Kagel, and Leonard Green. 1981. "Maximization Theory in Behavioral Psychology," *The Behavior and Brain Sciences*, 4-03, 371-388.

Recalde, Maria P., Arno Riedl, and Lise Vesterlund. 2014. "Error Prone Inference from Response Time: The Case of Intuitive Generosity." *Working Paper*.

Rosenthal, Robert. 1981. "Games of Perfect Information, Predatory Pricing, and the Chain Store." *Journal of Economic Theory* 25 (1): 92–100.

Roth, Alvin E. 1995. "Bargaining Experiments," in *Handbook of Experimental Economics*, edited by John H. Kagel and Alvin E. Roth, Chapter 4.

Roth, Benjamin, and Andrea Voskort. 2014. "Stereotypes and false consensus: How financial professionals predict risk preferences." *Journal of Economic Behavior & Organization*, forthcoming.

Shafir, Sharoni, Tom A. Waite, and Brian H. Smith. 2002. "Context-dependent violations of rational choice in honeybees (Apis mellifera) and gray jays (Perisoreus Canadensis)." *Behavioral Ecology and Sociobiology*, 51(2): 180-187.

Sutter, Matthias, Peter Martinsson, Francesco Feri, Katarina Nordblom, Martin G. Kocher, and Daniela Rützler. 2010. "Social Preferences in Childhood and Adolescence: A Large-Scale Experiment." IZA DP No. 5016 *Working Paper*.

Varian, Hal R. 2006. "Revealed Preference." Szenberg, M., L. Ramrattan, and A. A. Gottesman (eds.), *Samuelson Economics and the Twenty-First Century*, Oxford University Press, pp. 99–115.

von Gaudecker, Hans-Martin, Arthur van Soest, and Erik Wengström. 2012. "Experts in Experiments: How Selection Matters for Estimated Distributions of Risk Preferences," *Journal of Risk and Uncertainty*, 45(2), 159–190.

Tarr, David G. 1976. "Experiments in Token Economies: A Review of the Evidence Relating to Assumptions and Implications of Economic Theory." *Southern Economic Journal*, 43(2): 1136-1143.

Thaler, Richard H. 1997. "Giving Markets a Human Dimension." *Financial Times*, 6, June 16.

Wooders, John. 2010. "Does Experience Teach? Professionals and Minmax Play in the Lab." *Econometrica*. 78(3), 1143-1154.