Intelligent Design

The Relationship of Economic Theory to Experiments: Treatment driven Experiments

Muriel Niederle
Stanford University and *NBER*

August 13, 2010

When I interact with colleagues and friends who are venturing into experimental economics, either as they prepare for their own experiment, or aim for a critical view of experiments run by others, I often hear the question: "What is a good experiment?" My first reaction is to answer my empirically minded colleagues "Well, let me ask you: What is a good regression?" Clearly a good experiment (or regression) is one that allows testing for the main effect while controlling for other plausible alternatives. This helps ensure that the original hypothesis is reached for the right reasons and the initial theory is not wrongly confirmed.

However, there is an aspect of experimental design that is probably closer connected to theory than empirical work: As designers, we are responsible for the environment in which the data are generated. As such a good experimental design also needs to fulfill requirements one may impose on good theory papers: The environment should be such that it is easy to see what drives the result, and as simple as possible to make the point. The design has to provide a good environment for studying the questions at hand. Furthermore, ideally, the design (just like good theory) should be such that it seems plausible that the results could serve as prediction for behavior in other environments.

The design of experiments has some interplay with empirical methods as a good experimental design foreshadows how the data to be generated can be analyzed. As such, good design can often reduce the need for fancy econometrics, or, at times, allow econometrics to be more powerful. It also often implies that the experimental design has to take a stand on what it means to accept or reject an initial theory or hypothesis.

When deciding about a design, there are basically no fixed rules. The one exception probably being that economic experiments which use deception are really frowned upon, hard to run in many experimental labs, and often have a hard time to be published in economics journals.[1] Apart from that, however, anything goes. This may make it harder to design and evaluate a good experiment.

In this chapter of the book on methodology of experiments, I want to focus on the interplay between experimental design and the testing of hypotheses. This includes hypotheses that rely on theory, as well as some that are described less formally. Specifically, I want to show how in many cases intelligent design can provide direct tests, instead of having to rely on indirect evidence.

Section I shows a line of work where the test of the theory becomes more and more stringent, casting new doubts on the validity of the theory. In section II, I present several ways in which theory can be tested. First, I show how it might be useful to test assumptions directly rather than relying on econometrics. Then I discuss two methods to test whether a theory is responsible for the phenomenon, one can be called the "Two Way" design and the other the "Elimination" design. Finally, I discuss issues that are relevant when running a horse race among theories. In section III I show how the methods used when testing theory apply even when there is no detailed model. I show how experiments can be used in trying to understand the important channels that drive a phenomenon.

When talking about theory and the relation to experiments, an unavoidable question will be: When should a theory judged to be "good" in terms of relating to the experimental results? What should the criteria be in the first place? I will touch on these issues, as the chapter progresses.

I. BELIEF-BASED MODELS – A DIRECT TEST

The first example I want to delve in is how experiments have been used in the formulation and testing of new, less stringent theories on behavior in strategic games. I will present a series of approaches to test the theory, and show how new design made the test more and more direct, and may lead to some new questions as to how to think of those theories.

[1] For example, Econometrica states that "We understand that there may be a need for exceptions to the policy for confidentiality or other reasons. When this is the case, it must be clearly stated at the time of submission that certain data or other appendix materials cannot or will not be made available, and an explanation provided. Such exceptions require prior approval by the editors." http://www.econometricsociety.org/submissions.asp#Experimental, accessed 8/11/2010. For an overview on some of the arguments, see Hertwig and Ortmann (2008) and

It is well known that standard Bayesian Nash equilibrium makes very strong assumptions on rationality: First, players have to form beliefs about strategies of other players, and they have to best respond to these beliefs. Furthermore, in equilibrium these beliefs have to be correct. One simple modification is to relax the assumption that beliefs are correct, while maintaining that players form beliefs and best respond to them. A prominent version of such a modification is the k-level thinking model.

This model was created and became prominent by the results of the so called beauty contest or guessing game. In such a game, several players have to choose a number from some interval, e.g. 0 to 100. Furthermore, the player who is closest to, say, 2/3 of the mean of all players receives a prize. Obviously the Nash equilibrium is 0, but 0 is also never the winning guess. In a first experiment, Nagel (1995) showed that a large fraction of responses center around 2/3 of 50, and 2/3 of 2/3 of 50, see Figure 1 below.

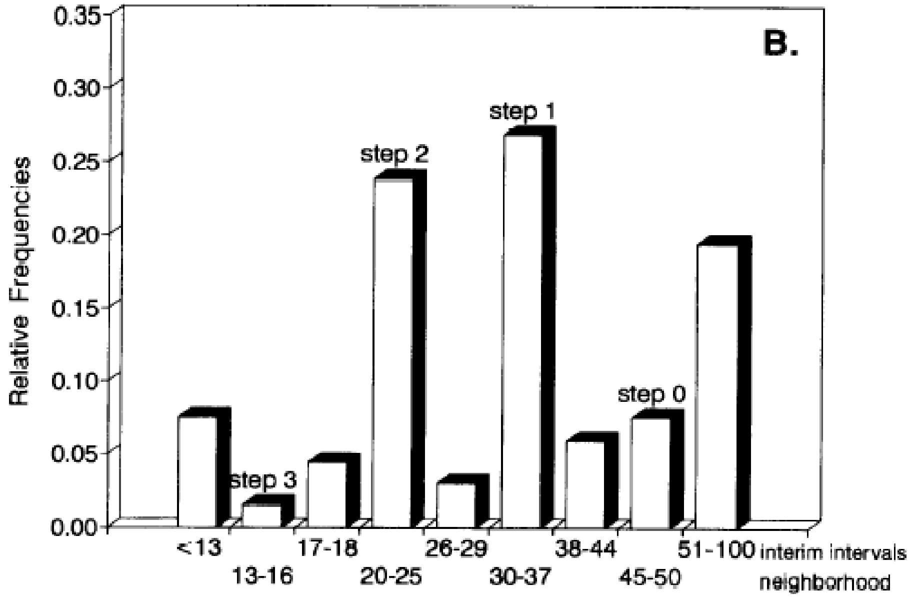

Figure I: Relative Frequencies of Choices in the First Period According to the k-level classification with a level 0 expected play of 0 (source: Nagel, 1995)

The behavior can be rationalized the following way. Those that play 2/3 of 50 may be participants that believe that other players simply pick numbers randomly, and hence the best response is to guess 2/3 of 50. Such players are called level 1 players (and correspond to step 1 in Figure 1), as they best respond to level 0 players, players that play non-strategically, and here are assumed to select numbers randomly. Participants that choose 2/3 of 2/3 of 50 best respond to a world that is fully comprised of level 1 players, and are hence called level 2 players. In general, for any given strategy of level 0, a level *k* player plays a best response to the belief that all other players are of level *k*-1 (see also Stahl and Wilson, 1995).[2]

I.A: ESTIMATING PARAMETER(S)

Many experiments on belief based models, such as Nagel (1995), show that *k*-level models are pretty good at organizing behavior observed in the lab and in the field.[3] The data are analyzed by estimating which fraction of players use strategies that are close to *k*-level strategies for various *k*'s and for an appropriately chosen level 0 play. Such participants are then classified as *k*-level players.[4] In principle such estimations allow a lot of freedom in the choice of a level 0 strategy, which may make fitting data to *k*-level play easy. The literature has in general put a restriction on level 0 play as random play with all actions chosen with the same probability (see e.g. the discussion in Camerer et al, 2004).[5] In the case of common value auctions with private information "naïve" play has also been proposed as a natural level zero play. Naïve play is often a strategy that depends in a naïve way on the agents' private information; often they are simply assumed to bid their signal (see Crawford and Iriberri, 2007).

The experiments in this category do not allow to more directly test whether the comparative static predictions of a level *k* model are correct, or, even more fundamentally, whether players actually form beliefs and best respond to them, which is the assumption at the heart of any belief based model, as well as, of course, the standard model.

In general, experiments that estimate parameters only allow for very limited conclusions. This may not be a problem when the aim is to estimate a specific parameter, such as, say, coefficients of risk aversion, while maintaining the assumption that players have well formed preferences over risky outcomes. It may not be a limitation either when finding a parameter that

[2] On a similar note, cognitive hierarchies (see Camerer et al, 2004) assumes that a level 0 player plays a random strategy, and a level *k* players best responds to a distribution of players of levels *k*-1 and lower, where the distribution is given by the restriction of a common Poisson distribution over all levels on the levels strictly below *k*.

[3] See Bosch-Domènech et al (2002) for a beauty contest game played in the field, and Ho, Camerer and Weigelt (1998), Costa-Gomes and Crawford (2006), Crawford and Iriberri (2007), Costa-Gomes, Crawford and Iriberri (2009), and also Stahl (1998).

[4] Some papers, such as Costa-Gomes and Crawford (2006) use a somewhat more stringent test by not only taking into account the actions of players, but also what information they used when making a decision. Specifically, parameters of the game were hidden and had to be actively uncovered by participants. The pattern of lookups can exclude misspecification of certain "random" strategies when the data needed to be a *k*-level player was not even uncovered. Camerer et al (2004) propose to fit data using a unique parameter that determines the distribution of types.

[5] However, there has not yet been too much of a debate about the predictive power of *k*-level models, and what behavior is easily excluded, even though such debates are active for other models of deviations of rational behavior, such as Quantal Response Equilibrium, see McKelvey and Palfrey (1995) and Haile Hortaçsu and Kosenok (2008).

suggests a theory is wrong. It is, however, more of an issue when trying to conclude that a specific theory is right. This may be especially the case when, at the same time, it is not accompanied by a discussion about what fraction of potential behavior is accommodated by the theory, and what fraction of potential behavior would reject the theory (where of course it may be hard to say what fraction of behavior can reasonably be expected to occur).

I.B: COMPARATIVE STATIC PREDICTIONS

Moving beyond experiments that (merely) serve to estimate a parameter, comparative static experiments on *k*-level thinking provide a somewhat more stringent test than fitting the data of an experiment.

For example, Costa-Gomes and Crawford (2006) test whether participants play differently in generalized two-player guessing games, depending on the strategy used by the opponent. In such a game, each player receives a lower limit and an upper limit in which to choose a number and a personal multiplier. The player is paid dependent on how close they are to their multiplier times the average of their own and their opponents guess. They want to assess whether participants respond to strategies of opponents. There are two kinds of ways of doing that. One is to elicit a subjects' beliefs about the strategy used by the opponent. This has the problem that gathering information about beliefs is hard, and may or may not influence actual play (see e.g. Costa-Gomes and Weizsaecker, 2008). A second approach is to manipulate the beliefs of a subject about the strategy of their opponent. This is the route chosen by Costa-Gomes and Crawford (2006).

Specifically, they have each participant play against a fixed strategy played by a computer that is explained to them in detail. For example, if the computer is programmed to play level 1, where level 0 is random play, the description says (see the web appendix to Costa-Gomes and Crawford, 2006, on http://dss.ucsd.edu/~vcrawfor/ accessed Jan 14, 2010).

> "The computer's rule is based on the assumption that you are equally likely to choose any guess from your lower limit to your upper limit, so that, on average, you guess halfway between your lower and upper limits. The computer's rule is to choose the guess that would earn it as many points as possible if you guess halfway between your lower and upper limits."

For level 2 the description is:

> "The computer's rule is based on the assumption that you assume that the computer is equally likely to choose any guess from its lower limit to its upper limit, so that, on average, it guesses halfway between its lower and upper limits.

> The computer's rule is to choose the guess that would earn it as many points as possible if you chose the guess that would earn you as many points as possible assuming that the computer guesses halfway between its lower and upper limits."

They say that participants are indeed able to best respond, to some extent.

Georganas, Healy and Weber (2009) also offer a more direct test of the hypothesis that participants change their level of play dependent on their opponent, where they manipulate the beliefs about the ability of the opponent. They have participants take an IQ test amidst several other tests, to construct a combined score. Participants played 10 games (generalized two player guessing games and other games), once against a randomly drawn opponent, once against the opponent with the highest combined score, and once against the opponent with the lowest combined score. The idea is that participants should adjust the level of strategic reasoning they assign to their opponent, depending on the opponents' general abilities. While they find that participants use somewhat higher level-*k* strategies against the opponent with the highest score compared to a random opponent, there is no significant downward shift in level when playing against the person with the lowest score. Furthermore, they find a lot of variation in a participant's level of sophistication across games. This is true not only at the absolute level, but also when comparing the relative sophistication of a subject compared to the depths of reasoning of all other participants.

Overall, comparative static tests of *k*-level thinking models seem to have limited success. Furthermore, across games, players cannot necessarily be described as being of a fixed level *k*. Similarly for the cognitive hierarchy models, estimations of the parameter of the Poisson distribution of level *k*'s yields differences across games, though Camerer et al (2004) summarize their paper as: "An average of 1.5 steps fits data from many games." Two remaining open questions are for what classes of games can a large fraction of "reasonable" behavior lead to results that are not consistent with the model and what games do allow for the *k*-level model to be falsified? A second issue is how to think of the statement "many games", one I will return to in the next section.

I.C: ARE THE UNDERLYING ASSUMPTIONS REALLY FULFILLED?

The most stringent interpretation of many papers on *k*-level models is that they indeed present a fair representation of the way in which many participants behave. That means participants form beliefs that other players are of some level *k* and best respond to them by playing a level *k*+1 strategy. The tests so far involve estimating whether a sizeable fraction of plays, and/or of players correspond to strategies that fall into the level *k* description (for an appropriately chosen level 0).

The more stringent test of a comparative static prediction that players adapt their play given the strategies of others has more mixed success.

How can we test directly whether players actually form beliefs and best respond to them? The strategy of manipulating beliefs instead of trying to assess them seems right. However, when using the approach of Costa-Gomez and Crawford (2006), one might worry that by describing, say, the level 1 strategy of the computer, participants may be "trained" in thinking about best responding, and hence may be more likely to become a level 2 player. Providing players with a description that includes best response behavior may change their own thought process. On the other hand, the strategy of Georganas, Healy and Weber (2009) rests upon players forming beliefs what strategies players use, when they are either the winner or loser of a quiz. While intelligence may be correlated with the depth of reasoning a player may be capable of, it need not be correlated with their beliefs about the level of reasoning used by their opponent. As such, clear predictions may not be that straightforward.

Ideally, we would like to know the beliefs players have about their opponents' strategy, without influencing their thought process, explaining strategies, or eliciting additional information. That is, we would like to provide a direct test whether players best respond to beliefs they have about strategies of their opponents. Hence, we need to design an environment where both the players and the experimental economist know the strategy of the players opponent, without providing the player with any additional information.

The environment we use to directly test *k*-level thinking models is overbidding in common value auctions, where participants in general fall prey to the winners curse. Overbidding, bidding above the Bayes Nash equilibrium, has been shown to be consistent with k-level thinking, and indeed is one of their success stories (see Crawford and Iriberri, 2007).[6]

In Ivanov, Levin, Niederle (2010), we propose a very simple two player common value auction. Players each receive a signal *x* between 0 and 10 (discrete values only), where the value of the item is the maximum of the two signals. Participants then bid in a second price sealed bid auction for the item. It is easy to see that bidding less than one's signal is a weakly dominated strategy, as the item is always worth at least one's own signal. Furthermore, a second round of iteration of weakly dominated strategies eliminates all bid functions that call for bids that are strictly greater than one's signal. In fact, bidding the signal is the unique symmetric Bayesian Nash equilibrium bid function.

[6] However, study applies to any belief-based explanation of the winners' curse. This includes, for example, cursed equilibrium (see Eyster and Rabin, 2005), and analogy-based expectation equilibrium (Jehiel, 2005, Jehiel and Koessler, 2008).

The general finding in common value auctions is the winners' curse, that is, participants bid above the Bayesian Nash Equilibrium which often results in the winner of the auction paying more than the value of the item. How can the k-level thinking model in this simple environment generate the winners' curse? Start with a level 0 player who bids randomly. Then the best response entails overbidding. The intuition is that in this case a level 1 player bids against a random number in a second price auction. Hence, the best response to random bids is to bid the expected value of the item which is in general strictly higher than the received signal, apart from the case when the signal is 10. Therefore, a level 1 player would be overbidding, compared to the symmetric Bayes Nash equilibrium. This implies that when two level 1 players bid against each other, we would confirm the standard result of a winners curse.

The nice feature of our setup is that a best response to a level 1 bidder (or any bidder who bids above their signal) is to bid the signal. In general, a best response to overbidding entails to certainly bid less than that. This allows for a very stark prediction of any model that includes best response behavior. Hence we have a very simple comparative static prediction if we manipulate the beliefs of at least some players, that their opponent is someone who bids above their signal. However, the goal is to achieve this without providing players with any information about possible strategies, so as not to affect their thought process.

In the experiment participants first play for 11 rounds the two player common value second price sealed bid auction against varying opponents, where each participant receives each signal $\{0,1,2,..,10\}$ exactly once. As such, we basically elicit the participants' bid function. As is common in the literature, to not distort the information participants have about the game, they receive no feedback, no information whether they won the auction, or what the item was actually worth.

Figure 2 shows for each signal the fraction of bids that fall into various categories, where we allow for small errors, and hence have bands of 0.25 around the signal. Note that bids $b(x) < x-0.25$ and $b(x) > 10.25$ can only be level 0 bids, bids $x+0.25 < b(x) \leq 10.25$ can be classified as level 1 bids (for a random level 0) and bids $b(x) \sim x$ as level 2 bids.

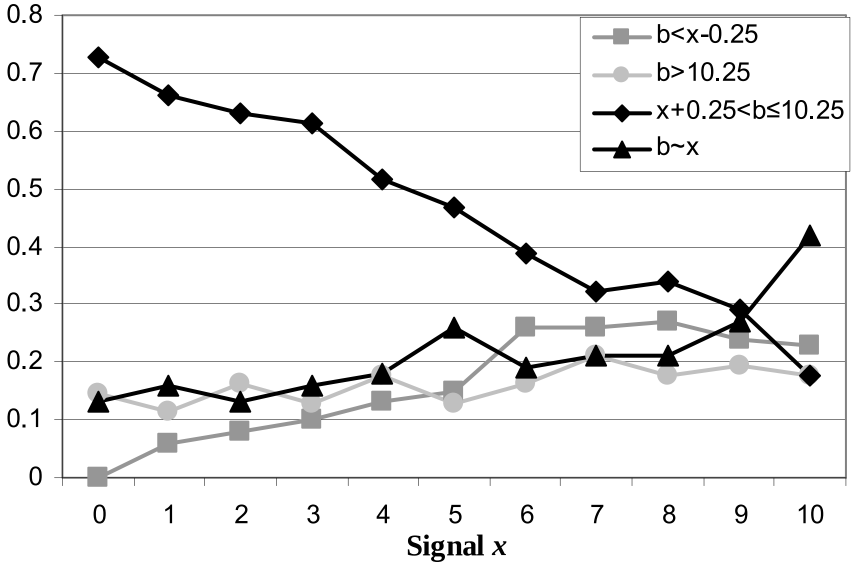

Figure II: For each signal, the fraction of bids that fall into various categories: "*b*<*x*-0.25" represents bids such that for a signal *x* the bid *b(x)* < *x*-0.25. Similarly for "*b*>10.25" and *x*+0.25<*b*≤10.25". "*b*~*x*" includes for each signal *x* bids in *x*-0.25 ≤ *b* ≤ *x*+0.25.

Note that for each signal *x*, all bids that are strictly below *x* are weakly dominated since the item is always worth at least *x*. Similarly, bids strictly above 10 are weakly dominated. Figure 2 shows that for each signal, the majority of bids can be classified as level 1 bids for a random level 0 player, as those participants place bids above their signal, but not above 10. A sizeable fraction of bids correspond to the symmetric Bayes Nash equilibrium, and can be thought of as level 2 bids.

	Phase I
Under-bidders	5
Signal-bidders	9
Over-bidders	25
10+: Above-10-bidders	10
Ind.: Indeterminate	13

Table 1: Subject classification in Phase I of the Baseline treatment.

Table I classifies bidders depending on where they place the majority of their bids (6 out of 11).[7] In Phase I, many bidders place the majority of their bids in a way consistent with *k*-level

[7] For the table, we use the following, slightly different classification: Underbid/Signal-bid/Overbid/Above-10-bid are bid of, respectively, (i) $b < x-0.25$, (ii) $x-0.25 \leq b \leq x+0.25$, (iii) $x+0.25 < b \leq 10$, and (iv) $b >$

thinking. While the 10 bidders who bid above 10, and the 5 underbidders, are clearly not rationalizable with k-level thinking, a total of 25 (40%) of participants can be classified as level 1 bidders, and 9 (15%) as level 2 (or higher) bidders.

So far, the experimental results provide another "success" story of *k*-level thinking. The majority of bidders and bids can be classified as level 1 or level 2. However, we now turn to a more direct test of the *k*-level model.

In the main treatment, Phase II of the experiment, participants play again 11 rounds of the two player common value second price sealed bid auction, where once more they receive each signal exactly once. This time, however, they play against a computer. The computer, upon receiving a signal y, will place the same bid that the participant placed in Phase I when receiving that signal y. That is, the computer uses the participants' own prior Phase I bid function. Hence, in Phase II, participants bid against their old bid function. This method allows us to have perfect control of a subjects' beliefs about the other players' strategy, since it is simply the subject's own past strategy. Furthermore, we achieved this without any interference, without providing any new information that may change the mental model of subjects. While the data of the treatment I present does not remind participants of their old bid function, we do so in another treatment, with virtually identical results.

Any best response model, and as such also the *k*-level model, predicts that players best respond to their own past strategy. Hence, a player of level k is expected to turn into a level k+1 player. This implies that overbidding should be reduced in Phase II compared to Phase I. Consider a participant who overbids in Phase I. Then, bidding the signal is a best response. Continuing to overbid but less so may or may not be a best response, depending on how much the participant overbid beforehand, and by how much the bid function is lowered. However, no change in behavior is clearly not a best response.

Figure III shows for each signal the fraction of bids that fall into level 1 play, and those that are close to bidding the symmetric equilibrium, both for Phase I (filled) and Phase II (hollow). The fraction of bids, both above and around the Bayesian Nash equilibrium, are virtually unchanged.

10, with the exception for $x = 10$, where a bid b of $9.75 \leq b \leq 10.25$ falls in category (ii), and only a bid $b > 10.25$ falls in category (iv).

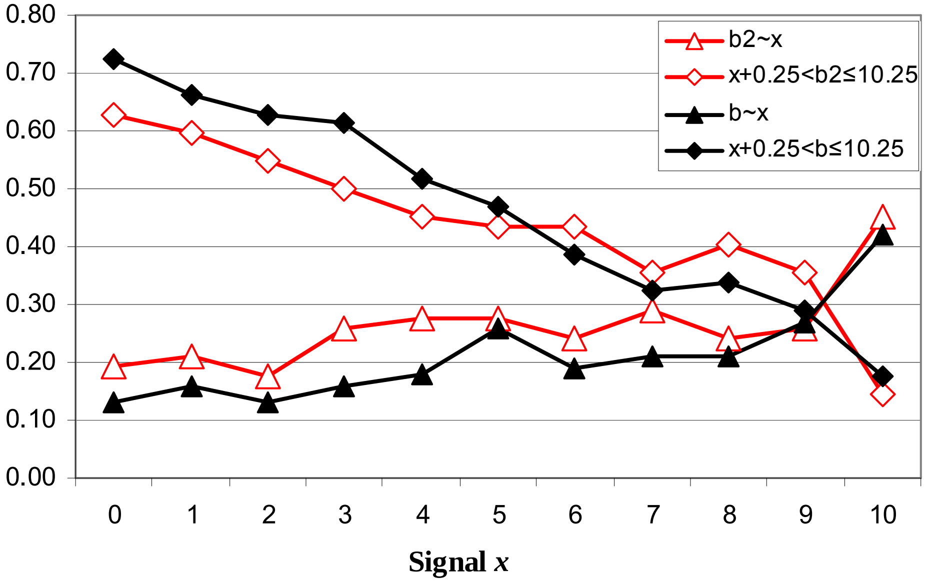

Figure III: For each signal, the fraction of bids that fall into various categories: "*b*<*x*-0.25" represents bids such that for a signal x the bid b(x) < x-0.25. Similarly for "b>10.25" and *x*+0.25<*b*≤10.25". "*b*~*x*" includes for each signal *x* bids in *x*-0.25 ≤ *b* ≤ *x*+0.25. *b* indicates a bid in Phase I and *b2* a bid in Phase II.

Furthermore, we characterize bidders depending on where they placed the majority of their bids, and determine whether bidders of various types changed their classification.

Table 2 shows that overbidders mostly remain overbidders (56%), only a minority (24%) turn into underbidders or signal bidders. For subjects who are overbidders in parts I and II, we find that only 23% of bids in Phase II are best-responses to Phase I behavior. By not behaving optimally in part II, these subjects are on average foregoing more than 20% of the earnings an average subject made in the course of the experiment.

	Under	Signal	Over	10+	Ind.	**Phase I**
Under-bidders	2	0	2	1	0	**5**
Signal-bidders	0	5	3	1	0	**9**
Over-bidders	1	5	14	1	4	**25**
10+: Above-10-bidders	2	1	1	6	0	**10**
Ind.: Indeterminate	2	2	3	5	1	**13**
Phase II	**7**	**13**	**23**	**14**	**5**	

Table 2: Subject classification in parts I and II of the Baseline treatment, depending on how they placed the majority (6 out of 11) of their bids.

We investigated whether the winners' curse in common value auctions can be rationalized using belief-based models such as *k*-level thinking. We achieved a direct test of this hypothesis by comparing behavior in environments in which overbidding can be rationalized by overbidding (such as in Phase I play) and in environments where it cannot (such as in Phase II play for participants who were overbidders in Phase I). The results of the experiment are in general bad news for any theory that relies on best response behavior, especially in complicated environments such as second price sealed bid common value auctions. This is despite the fact that our environment seems much easier than some other environments used in common value auction experiments.

This poses questions as to what we think a theory should accomplish. I will restrict attention here to musing about the role of theory in understanding behavior in experiments. Certainly, theory has a big role to play, when, e.g. deciding to give advice, whisper into the ears of princes, and, for example, perform real design (see Roth, 2008). Should a theory be merely a tale, providing insights, but otherwise, like a fable, clearly wrong, clearly leading to absurd conclusions or predictions, but, once more like a fable, with a kernel of truth? This view is certainly not unheard of (see Rubinstein, 2006).

Alternatively, one might want a theory to fit data very well, and so, a theory might be judged by how well it reconciles with a large set of data. This is largely the approach taken by proponents of new belief-based models. In general, they fit the data of the experiments that are considered better than standard theory, and as such have been deemed superior in providing a model of how agents behave in such environments.

Finally, one might want a theory to provide good (i.e. accurate) predictions, either in general, or at least in terms of comparative statics, to guide us in predicting how behavior would change, if we were to make certain changes to a game. As such, a good theory may also provide insights as to what are the important parameters of a game that are likely to affect behavior in the first place. Note that in principle one theory might be a better able to fit data on any given experiment (though with different parameters), while another theory might be better at making predictions out of sample.

In this section, the model of *k*-level thinking seems to be less promising as a model that is absolutely right, that makes good comparative static predictions when changing the game. Though maybe, in this game, participants are so much at a loss, that the model simply does not apply, that maybe, using it to explain behavior in common value auctions was simply too ambitious a goal? It still remains a question, whether the model might be able to predict behavior in other instances. As such, this goes back to the question whether we should be content when a model can fit the data of multiple experiments (though maybe with varying parameters, or varying proportions of level 1 and level 2 players). Or do we require the model to make predictions out of sample? This can include a good fit on new games, which is the approach that has mostly been taken by level-*k* proponents so far. It can also include predicting comparative static behavior better than other models (which has received less attention so far).

Finally, assume the model cannot predict out of sample behavior, by trying to fit behavior that is obtained when certain parameters of the game change (such as manipulating the beliefs about the strategy of the opponent). It might still be the case, that the description of the data in itself might be valuable. For instance, it could be that the fraction of level 2 players (or the specific *k* that fits the cognitive hierarchy model to the data of a specific game) may be a good predictor for, e.g., how "difficult" a game is. It could be a decent proxy for the complexity of a game that could be used to further our understanding on what it is that makes some games harder than others.

Later in the next section, I will revisit the question of how to think of theories in the face of experimental evidence.

II. Experimental Design and Hypothesis Testing

The first example showed how design can allow for a direct test, instead of simply having to indirectly estimate whether behavior confirms to the model at hand. Note that the design reduced the need for econometrics. In this chapter, I want to provide a few more examples of that sort, where intelligent design changed the problem so that the hypothesis could be attacked directly. However, I also want to show how sometimes intelligent design is needed in order to be able to provide an environment in which a hypothesis can be tested in the first place. The following can be thought of as a toolbox for designing experiments.

II.A. Testing for Selection

In this first section I want to elaborate on a design that allows for a comparison between the power of an intelligent design, and the result of using standard econometric techniques. Once more the environment is one of overbidding in common value auctions, this time first price auctions.

While overbidding (bidding above the risk neutral Bayesian Nash equilibrium) is the standard behavior in common value auctions, later rounds of experiments show fewer instances of participants falling prey to the winners' curse compared to earlier rounds. Furthermore, experienced participants, participants who return sometimes weeks later, suffer from the winners' curse less than those who participate for the first time. There are two potential reasons for why participants seem to learn to use better strategies. One is that less able bidders may simply go bankrupt in the course of the experiment (having suffered repeatedly from the winners' curse) and are barred from further participation, as they have no more earnings they can lose. Furthermore, when experience is measured across sessions taking place on separate days, participants who overbid a lot, and made less money, may not return at the same rate for subsequent experimental sessions. Hence, one reason for better performances in later rounds could be a pure selection effect. Second, it could be that bidders indeed learn to avoid the winner's curse, where learning can occur in the first session, between sessions and during the second session.

One aim of Casari, Ham and Kagel (2007) is to design an experiment that can directly measure the effects of selection, compared to learning. To reduce the effect of selection, they have treatments in which some participants have higher initial cash balances than others, and throughout the experiment also receive windfall gains (via a lottery), thereby affecting the probability with which a subject will go bankrupt during the course of the experiment. To study the importance of selection in accounting for the fact that experienced participants perform better, Casari, Ham and Kagel (2007) vary the show up fees for the second week, and for some participants even hold half the earnings of the first week in escrow, to ensure a high return rate for a subset of participants.[8]

The control treatment employs standard procedures: participants receive a show up fee of $5, and an initial cash balance of $10. All subjects were invited back to week 2, where they received the same show up fee and cash balance once more. In the bonus treatment starting cash balances were $10 for half the participants and $15 for the other half. Furthermore, after each auction, each active bidder participated in a lottery that paid $0.5 with 50%. In addition, a show up fee of $20 was paid only after completing week 2's session, with 50% of the earnings of the first week held in escrow as well. A third random treatment, was similar to the bonus treatment only that participants all received a show up fee of $5 in week 1, and either $5 or $15 in week 2.

Within the first session, the changes in design indeed affect bankruptcy rates they range from 46.3% to 20.3%. Similarly, the rates of participants returning for a second experimental session vary from 96% to 60%. A first clear indication of selection of participants is to consider among the subjects that went bankrupt in the first session, how many return to a second session. In the baseline treatment, this is only 47.7%, while it is 88% in the treatments that made bankruptcy harder in the first place! Nonetheless, in all treatments the authors find that

[8] In their common value auctions, the value of the item x is a random uniform draw from [$50, $950], where each of the six bidders receives a private signal y, drawn independently from [x-$15, x+$15]. Because of boundaries, when attention is restricted to x in [65, 935], the bid factor (the amount to be deducted from the signal) of the RNNE is about 15, which is close to the loss-free strategy (where bidders can ensure never to lose money). The bid factor for the break-even strategy (the strategy that yields zero expected profits with occasional looses) is about 10.71.

participants learn in that their behavior is closer to the RNNE, albeit to various degrees. It seems that both market selection and individual learning is responsible for improved outcomes of experienced bidders.

The authors also show that standard econometric techniques fail to provide evidence of unobserved heterogeneity in the bidding behavior of participants, and fail to detect any selection effects.

The paper shows how intelligent design can address selection (or potentially other econometric issues) directly, instead of having to rely on sophisticated econometric techniques, which may fail to find any evidence (which could be due to the fact that the samples of experimental economists are typically smaller than of labor economists).

II.B. TWO WAY DESIGN

Overbidding, bidding above the risk neutral Nash equilibrium (RNNE) by participants is not only a regular phenomenon for common value auctions, but also in private value first price auctions. Early work has attributed that to risk aversion (see e.g. Cox, Smith and Walker 1988), which calls for bids above the RNNE, closer to one's valuation. That literature has spawned a serious critique, namely that behavior of participants in first price auctions may be hard to take seriously, as the change in expected payoff is quite flat around bid functions that correspond to the risk neutral Nash Equilibrium (see Harrison 1989). This spawned a very lively debate.[9]

In a very clever design, Kagel and Levin (1993) want to show that in general, overbidding in first price auction may not necessarily be attributable to either risk aversion, or the possibility that participants simply don't care about the bids they place.

One way to address this question would be to try to estimate each participant's level of risk aversion, and correlate it with the bids they place. Note that we would still need to heavily rely on the Nash equilibrium model, implying that subjects form correct beliefs about the degrees of risk aversion of other participants and so on.

Kagel and Levin (1993) present a much more direct, simpler and elegant solution. Specifically, they have participants bid not only in first price but also in third price auctions, auctions in which the highest bidder receives the object for the third price. In this case, the RNNE calls for bidding *above* one's valuation, and risk aversion calls for bids *below* the RNNE. Both of these are in contrast to behavior expected in the first price auction, where RNNE calls for bids below one's valuation, and risk aversion for bids above the RNNE.

[9] See Friedman (1992), Kagel and Roth (1992), Cox, Smith and Walker (1992), Merlo and Schotter (1992), and Harrison (1992).

They find that bidders bid above the RNNE in both auctions formats, which makes it more unlikely that the overbidding in the first price auction can be attributed to risk aversion (see also Kagel 1995 and Kagel and Levin, forthcoming.)

In this subsection we showed how an intelligent design can cast doubts on a prominent theory for a phenomenon. This was achieved by changing the environment such that the prominent theory (risk aversion) would now make opposite predictions to other theories that may account for the initial phenomenon, such as “wanting to win” for the “bidding above the risk neutral Nash equilibrium in first price auctions” phenomenon.

To elaborate on the use of such a two way design, I want to provide maybe one of the earliest examples of such a design from the book of Judges of the Old Testament, Chapter 6 (by courtesy of Al Roth’s experimental economics class).

The story begins with the Israelites, turned away from God after 40 years of peace brought by Deborah's victory over Canaan, being attacked by the neighboring Medeanites. God chose Gideon, a young man from an otherwise unremarkable clan from the tribe of Manasseh, to free the people of Israel and to condemn their worship of idols. Gideon, to convince himself, that the voice he hears is indeed the voice of God, asks for a test:

> And Gideon said to God: 'If You will save Israel by my hand, as You have said, look, I will put a fleece of wool on the threshing-floor; if there be dew on the fleece only, and it be dry upon all the ground, then shall I know that You will save Israel by my hand, as You have said.'
> And it was so; for he rose up early on the next day, and pressed the fleece together, and wrung dew out of the fleece, a bowlful of water.

So far, this makes Gideon merely an empirically minded person, not a good designer of experiments. Gideon, however, realized that there could be an alternative explanation for the main result. It could be that this is simply the way it is, after all, he probably wasn’t used to leaving a fleece outside. His design so far only allows for a test to see whether the result is due to God almighty, or merely the way cold nights interact with a fleece left outside. In his next design, he removes the “natural” explanation, and tests whether it is God’s doing, in which case, he can reverse the result that may be due to nature only.

> And Gideon said to God: 'Do not be angry with me, and I will speak just this once: let me try just once more, I ask You, with the fleece; *let it now be dry only upon the fleece, and upon all the ground let there be dew.'*
> And God did so that night; for it was dry upon the fleece only, and there was dew on all the ground.

The experiment also allows for predictions outside of the experimental design: Gideon raised the army which indeed defeated the Medeanites.

This is a prime example of an intelligent design: Change the environment such that the hypothesis of choice (God) would reverse the result, while the alternative (Nature) would leave the outcome unchanged. While many designs do not necessarily have comparative statics that work that beautifully, these are often useful in convincing the audience that the results are indeed driven by the specific hypothesis at hand.

II.C. ELIMINATION DESIGN: TESTING A THEORY BY ELIMINATING ITS APPLICABILITY

Apart from the two way design, there is another prominent method which I call the "Elimination" design, to cast doubt on the applicability of a theory to a particular phenomenon. Change the environment in a way the problem mostly stays the same, while, however, eliminating the conditions that allow the theory at hand to account for the phenomenon. That is, instead of testing the theory directly, examine to what extent similar behavior can be found in environments in which the model has no bite. If behavior remains unchanged, at least it implies that other factors may be at work as well. I will present two examples in detail that make that point, both of which are in environments which I already presented.

The first example concerns the guessing game which was introduced in Section I, which was a breeding ground for new theory, such as *k*-level thinking. The next experiment takes a phenomenon, such as players guessing a number not equal to zero in a guessing game, and changes the environment in a way to eliminate rationalizations of such behavior due to *k*-level models.

Grosskopf and Nagel (2008) change the guessing game to have only two, instead of three or more players. The rule is that the person who is closest to two thirds of the average wins. With two players this translates to the winner being the one who chooses the smaller number. That is with two players, guessing 0 is a dominant strategy. Hence any model that relies on players having non-equilibrium beliefs about the opponents to justify non-equilibrium behavior has no bite, as there exists a unique dominant strategy.

Grosskopf and Nagel (2008) find that among students, the guesses are virtually identical to the guesses made when the number of players was larger than 2, specifically, the instances of 0 were the same about 10% in both cases. While professionals (game theorists) are more likely to choose 0 when there are two rather than three or more players, zero is still chosen only by about 37%.

The findings of Grosskopf and Nagel (2008) cast serious doubt on the need for *k*-level models to explain guesses above 0, since such guesses are common even when 0 is a dominant strategy. However, there remains the possibility, that, when there are dominant strategies, other biases become important, or that participants were simply confused, and did not take into account that there were only two players.

A second example of an elimination design I want to provide is given in Ivanov, Levin and Niederle (2010). Remember that one way to justify overbidding, bidding above the signal in the simple two player second price sealed bid common value auction, is to assume that the opponent sometimes bids below the signal. In one treatment, we eliminate the possibility for participants to place a bid that is strictly below the signal (we call it the MinBid treatment).

We can compare the proportion of bids, and the proportion of bidders who can be classified as bidding above the signal in Phase I of the Baseline treatment (the treatment which was discussed at length in section I.C.), where participants play against a random person, and could place any bids they wanted below 1000000. If the main reason for overbidding in the Baseline treatment is that players believe that others may be playing randomly and sometimes underbid, then we would expect a large reduction of overbids in the MinBid treatment. However we find the opposite, an increase in the fraction of overbids, from about 40% to 60%. Overbidding is probably more frequent in the MinBid treatment because underbidding is impossible so that all bids are distributed in three, rather than four, categories. Given this, the frequencies of overbidding seem quite comparable.

The findings of Ivanov, Levin and Niederle (2010) cast doubt on belief-based models being the driving factor behind overbidding in common value auctions. While that third treatment, the MinBid treatment, eliminates any explanatory power of the theory, it could still be, potentially, that other biases become important, that are of a similar magnitude.

To summarize, one way to weaken the hypothesis that a theory can account for a phenomenon is to remove the applicability of the theory and show that the underlying phenomenon is virtually unchanged. However, it is still potentially possible that in this slightly changed environment there are other forces at work that yield similar results. It does not directly exclude that the theory has (at least also) some explanatory power for the phenomenon at hand. As such, providing a direct test may prove more fruitful to convince proponents of the theory than such indirect, albeit quite elegant, tests.

II.D. RUNNING A HORSE RACE AMONG THEORIES

Finally, I want to come to a quite popular method when comparing the predictive power of different theories. Before, I argued that theories may be valuable both in making point predictions or, alternatively, in predicting comparative statics due to changes in the environment. When it comes to papers that run a horse race among theories, the valuation is in general driven by how well each theory is able to fit the data in a fixed set of games, almost no attention is given to comparative static predictions.

In many papers such horse races involve a preferred theory (mostly a new theory of the authors of the study), and some more “standard” or established theories. Often, authors would pick a few games, have subjects play those games, and then compare the results across these games. It is not uncommon to run regressions often showing how a certain favorite theory does better than other theories, when giving equal weight to each of the games selected.

Here are my two biggest concerns with papers of that sort: The first is a lack of a deep discussion of how the games have been chosen. Often, such discussion is short, and not precise. This makes it hard to interpret any econometrics based on those games: What does it mean that the author can pick, say, 10 games in which the authors’ model does better than the other models? What does it even mean to run such a regression? Clearly, few people would be impressed if the authors tried many many more games and then selected a few such that the authors’ preferred theory wins the horse race. While such blatant abuse is probably rare, often authors may simply more easily think of games that confirm their intuition.[10] It may be hard to assume that the choice of games is not in some way biased. Of course, a potentially equally big problem is if the winning theory is designed after the fact, as I hope to make clear in the next paragraph.

This critique goes often hand in hand with my second problem with such work: One has to be careful that new theories are not, what could be called “toothbrush” theories: Theories one wants to use only on one’s own data (just like a toothbrush is for a single user only). What it means is that many papers take the form: “*My* theory works better on *my* data than *your* theory works on *my* data”. As such, it is clear why that is often not that impressive…

A paper that runs a horse race among theories has to first decide what success means. Is success really better predicting behavior in a few very specific games? Or is success rather predicting behavior in a given class of games?

In this chapter I want to advocate for the latter which is nicely described in Erev, Roth, Slonim, and Barron (2007). They test the predictive power of various models in two player zero sum games, including among others standard Nash predictions and learning models. In order to do that, they have participants play in a set of zero sum games with a unique mixed strategy Nash equilibrium. However, since the models are supposed to make a good prediction on that whole class of games, the authors did not pick the games themselves, but rather chose them randomly.

[10] See Roth (1994) about what constitutes an experiment, and his argument to keep up the integrity about how much “search” went on in terms of how many (unreported) trials or treatments were run in order to find a specific result.

This is a good strategy, unless there is a very good reason to focus only on a specific set of games.

That it may be easy to even fool oneself (taking the view that authors hopefully did not want to fool only others) when choosing games can be seen in the literature that precedes Erev, Roth, Slonim and Barron (2007) and Erev and Roth (1998). Previous papers can roughly be put in two groups: Papers that concluded that Nash is a good description of behavior, and papers that did not. As it seems, however, these two groups can also be characterized the following way: Proponents of Nash tended to pick two player zero sum games where the mixed strategy equilibrium resulted in relative equal payoffs for both players. The others tended to pick games where this was not the case. As no paper actually made that point precise, I imagine it was not due to earlier authors trying to fool others; their intuition may simply have led them to pick games that confirm their initial hypothesis.

Most recently, there have been several contests, whose goal was to find a model that predicts well in a class of games, where the exact games to be played will be randomly drawn (see Erev, Ert and Roth, 2010 and Erev et al, 2010).

III. WHAT CHANNELS DRIVE A RESULT? TREATMENT DRIVEN EXPERIMENTS

So far I have focused on experiments that test theories with a very precise underlying model. The experiments in this section go mostly beyond simple parameter testing or testing the comparative static of a well developed theory that makes precise predictions. The aim here is to go deeper into understanding the mechanism behind an initial finding. What are the channels that drive the initial result? Note that devising treatments to better understand the channels behind a specific result can of course be done, whether the initial finding is well grounded in theory or not. Similarly, the possible channels may be given by economic theory, though they may also be derived by finding from other disciplines, such as psychology.

The focus of this chapter is to describe the power of experiments in terms of understanding the channels that drive a result. This allows us to turn on and off various channels, which allows us to measure their impact, and really understand what drives a phenomenon. It is this powerful ability to control the environment that makes experiments so useful, especially when trying to understand what mechanism is responsible for the initial result, what are the key channels.

In this section, I will focus on the topic of whether women shy away from competition, and possible explaining factors. While Niederle and Vesterlund (2007), do not test a specific theory, we can ask, theoretically, what are potential economic reasons for (gender) differences in the decision to enter a tournament. Since we use a within subjects design, we need to have all controls ready before we start running the experiment (which is what makes such designs harder to implement). The advantage is that it will be easier to discern, how much various factors contribute to the decision to enter a tournament, as opposed to merely being able to say that they have some impact.

We aim to test whether women and men enter tournaments at the same rate, where their outside option is to perform in a non-competitive environment. Clearly, one reason for gender differences is choices of compensation scheme can be gender differences in performance. Gneezy, Niederle and Rustichini (2003) show that average performances of women and men in tournaments may be very different from each other, even when performances under a piece rate incentive scheme were very similar. In fact, we show a significant change in gender differences in performance across incentive schemes. Hence, in NV we aim for a task in which performance does not vary too much with the incentive scheme. Furthermore, all performances, both in competitive and non-competitive environments are assessed when determining the choices of women and men. Beyond that, what are possible contributing factors to gender differences in choice of incentive scheme?

Explanation 1: Men enter the tournament more than women because they like to compete.

This will be the main hypothesis, so, we have to design the experiment such that we can rule out other explanations.

Explanation 2: Men enter the tournament more than women because they are more overconfident. Psychologists and economists often find that while both men and women are overconfident about their relative performance, men tend to be more overconfident than women (e.g., Lichtenstein, Fischhoff, and Phillips (1982), Beyer (1990), Beyer and Bowden (1997) and Mobius et al (2010)).

Explanation 3: Men enter the tournament more than women because they are less risk averse. Since tournaments involve uncertain payoffs, potential gender differences in risk attitudes may affect the choice of compensation scheme.[11]

[11] Eckel and Grossman (2002) and Croson and Gneezy, (2009), summarize the experimental literature in economics and conclude that women exhibit greater risk aversion in choices. A summary of the psychology literature is presented by Byrnes, Miller, and Shafer (1999). They provide a meta-analysis of 150 risk experiments and demonstrate that while women in some situations are significantly more averse to risk, many studies find no gender difference.

Explanation 4: Men enter the tournament more than women because they are less averse to feedback. One consequence of entering the tournament is that the individual will receive feedback on relative performance.[12]

The experiment has groups of 2 women and 2 men seated in rows, and we point out that participants were grouped with grouped with the other people in their row. While participants could see each other, we never discuss gender during the experiment.[13] The task of our experiment is to add up sets of five two-digit numbers for five minutes, where the score is the number of correct answers. After each problem, participants learn the number of correct and wrong answers so far, and whether the last answer was correct. Participants do not receive any feedback about relative performance (e.g.. whether they won a tournament) until the end of the experiment.

The experiment has four tasks, where one of which will be randomly chosen for payment at the end.

Task 1—Piece Rate: Participants are given the five-minute addition task. If Task 1 is randomly selected for payment, they receive 50 cents per correct answer.

Task 2—Tournament: Participants are given the five-minute addition task. If Task 2 is randomly selected for payment, the participant who solves the largest number of correct problems in the group receives $2 per correct answer while the other participants receive no payment (in case of ties the winner is chosen randomly among the high scorers).[14]

In the third task participants once again perform the five-minute addition task but this time select which of the two compensation schemes they want to apply to their future performance, a piece rate or a tournament. The choice between the piece rate and the tournament should allow predicting money maximizing choices. Hence the choice must be independent of the subjects' beliefs about other players' choices which would otherwise enter the maximization problem and hence make it theoretically difficult to make money-maximizing predictions. This implies that the choice of each participant cannot influence other participants' payoffs.

Task 3 – Choice: A participant who chooses piece rate receives 50 cents for each correctly solved problem. A participant who chooses tournament has the new task-3 performance

[12] For example, Mobius et al (2010) explicitly ask participants about their willingness to pay (or get compensated for) receiving information about their performance in an IQ-like quiz. We find that men are significantly less averse to receiving feedback than women.

[13] We did not want to trigger any demand effects or psychological biases such as priming by pointing out that we study gender.

[14] On a technical note, by paying the tournament winner by performance rather than a fixed prize, we avoid providing information about winning performances, or distorting incentives for very high performing individuals.

compared to previous task-2 tournament performance of the other participants in his or her group. If the participant has the highest performance she or he receives $2 for each correct answer, otherwise she or he receive no payment. This way a choice of tournament implies that a participants' performance will be compared to the performance of other participants in a tournament.

Furthermore, since a participant's choice does not affect the payment of any other participant we can rule out the possibility that women may shy away from competition because by winning the tournament they impose a negative externality on others.[15]

Task 4—Choice of Compensation Scheme for Past Piece-Rate Performance: Participants decide between the piece rate and tournament incentive scheme for their task 1 piece rate performance, where a tournament choice results in a payment only if the participant had the highest task 1 piece rate performance in their group. This choice mimics the task 3 choice, while eliminating any tournament performance. Specifically, this choice, like the choice in task 3, requires that participants decide between a certain payment scheme (piece rate) and an uncertain payment scheme (tournament), receiving feedback whether their performance was the highest or not (tournament) or not receiving such information (piece rate). Like in task 3, participants have to base their decisions on their beliefs about their relative performance in their group.

As such, this treatment will provide some insight to what extent choices between incentive schemes in task 3 are affected by factors that are also present in task 4 compared to the unique factor that is missing in the task 4 choice, namely the desire or willingness to perform under a competitive incentive scheme.

Finally, participants provide information about their rank among the four players in each group both in the task 1 piece rate and the task 2 tournament performance (where a correct rank is rewarded by $1).

We find that women and men perform very similarly in both the piece rate scheme and the tournament scheme. The average performance in the piece rate is 10.15 for women, and 10.68 for men, and in the tournament is 11.8 and 12.1 respectively. There is no gender difference in performance. The increase in performance between the piece rate and the tournament seems to be due more to learning how to perform this task, rather than increased effort in the tournament.[16] Of the 20 groups, 11 are won by women, 9 by men, and men and women with the same performance

[15] There is a large literature on the debate whether women are more altruistic than men and hence may be more or less worried about imposing a negative externality on other participants (See Croson and Gneezy 2009, and Andreoni and Vesterlund 2001).

[16] This is supported by the fact that changes in performance between task 2 and task 3 are independent of the chosen incentive scheme in task 3. Note that this does not imply that participants do never provide effort, rather it appears their baseline effort is already quite high.

have the same probability of winning the tournament. Given the past tournament performance, 30% of women and 30% of men have higher earnings from a tournament scheme, which increases to 40% and 45% respectively when we add participants who are basically indifferent. However, in the experiment, 35% of women and 73% of men enter the tournament (a significant difference).

Figure IVa shows for each task 2 tournament performance quartile the proportion of participants who enter the tournament. Men have a higher chance to enter the tournament for any performance level.[17]

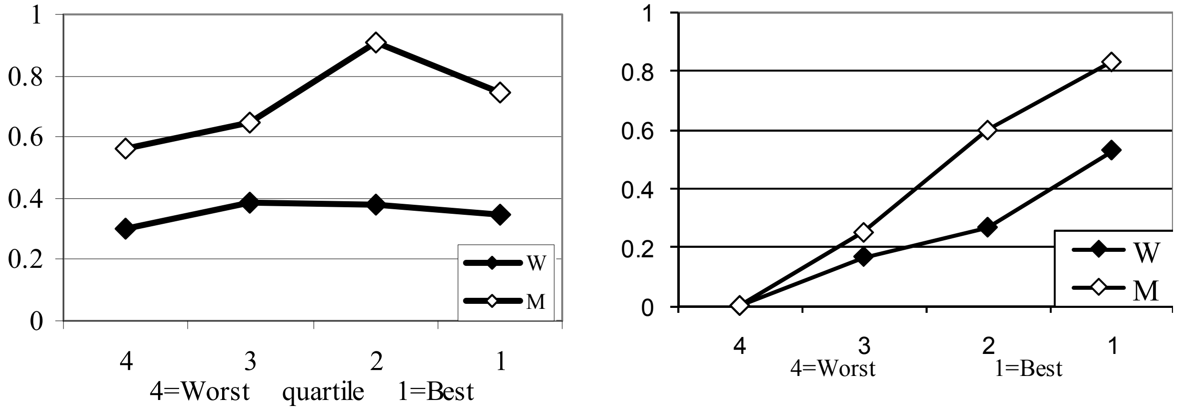

Figure IV: Proportion Selecting Tournament (Niederle and Vesterlund, 2007)

Panel A: Conditional on Initial Tournament Performance Quartile

Panel B: Conditional on Believed Performance Rank in Initial Tournament

One driving factor could be that women and men differ in their beliefs about their relative performance in the tournament (explanation 2). In the experiment 30 out of 40 men (75%!) believe that they were the best in their group of 4 (most of them were obviously wrong), men are highly overconfident. While women are also overconfident, 17 (40%) believe they had the highest performance, men are significantly more overconfident than women. Can this gender difference account for differences in tournament entry?

Figure IVb shows the proportion of participants that enter the tournament as a function of their guessed rank in the task 2 tournament. Beliefs have a significant impact on the decision to enter the tournament, but gender differences remain even after controlling for beliefs, which

[17] Similar results are obtained when we consider the performance after the entry decision, rather than the one before the entry decision.

account for about 30% of the initial gender gap in tournament entry (a result confirmed by regressions).

To study the impact of risk and feedback aversion on the decision to enter a tournament, we first study the decision in task 4 where participants decide whether to submit the task 1 piece rate performance to a piece rate or a tournament payment scheme. In this case the participants' actual performance, and their beliefs about relative performance can largely account for choices of women and men, the remaining gender gap in choices is economically small and not significant. That is gender differences do not follow the pattern found when choosing whether to enter a tournament and then perform. We studied a choice that mimics the decision of task 3, where participants decide whether to enter a tournament and then perform, only now, in task 4, participants did not have to perform anymore. Rather the payment was based on their past piece rate performance. In this case, we eliminate the need for an upcoming tournament performance, the decisions of women and men can be entirely accounted for by their actual performance and their beliefs about their relative performance. This already casts doubt that risk and feedback aversion (explanations 3 and 4) are major factors in determining gender differences in choosing to enter tournament. Furthermore, a regression on the task 3 decision of women and men to enter a tournament and then perform finds substantial gender differences, even when controlling for performance and beliefs and the choice in task 4.

In terms of money maximizing choices, high performing women enter the tournament too little and low performing men too much (though by design their losses are smaller, as payments are dependent on performance). The result is that few women enter the competition and few women win the competition.

Experiments were useful in showing this gender difference, as they allowed for controls that would be hard to come by with labor data: Apart from being able to control performances in both environments, we could also ensure participants that there are no aspects of discrimination whatsoever, their payments only depended on their decisions and performances of others, women were not treated differently than men. Furthermore, the experiments last less than 90 minutes; as such any concerns about raising children are clearly not an issue. Meanwhile, many other experiments have replicated the basic result.[18]

Some final comments about the design choices in Niederle and Vesterlund (2007). For example, when trying to understand the impact of risk aversion, or aversion to receive information about whether one's performance was the best or not, we could have chosen different ways. For example, we could have tried to explicitly measure risk aversion, by having

[18] For an overview see e.g. Niederle and Vesterlund (2010).

participants play various gambles, and measure their willingness to accept or reject those gambles. There are two comments to that approach.

The first concerns the actual risks or lotteries involved in the choices of participants. For example, for participants who have 14 or more correct answers the chance of winning the tournament is 47% or higher. If participants maintain the performance after their choice of compensation scheme, the decision to enter the tournament becomes a gamble of receiving, per correct answer, either $2 with a probability of 47% (or more), or receiving 50 cents for sure. Hence, a gamble of a 47% chance of $28 (i.e., an expected value of $13), versus a sure gain of $7. Of all participants who solve 14 problems or more, 8/12 of the women and 3/12 of the men do not take this gamble.[19] Similarly, for participants who have 11 or fewer correct answers the chance of winning the tournament is 5.6% or less. Thus entering the tournament means receiving $2 per correct answer with a probability of 5.6% (or less) versus receiving 50 cents for sure. For all participants who solve 11 correct answers this is a choice between a 5.6% chance of winning $22 (i.e., an expected value of $1.23) compared to receiving $5.5 for sure. Of the men who solve 11 problems or less 11/18 take this gamble while only 5/17 women do.[20] I am not aware of any studies that find such extreme gender differences in risk aversion. Furthermore if risk aversion is the most important explanation for the gender gap in tournament entry, men should not enter the tournament with a higher probability than women for all performance levels, but rather the entry decision of women should be shifted to the right of that of men.

Second, even if we found that more risk averse participants enter the tournaments at a lower propensity, it is not clear that risk aversion may indeed be the explaining factor. Similarly, if we found that risk aversion on its own does not reduce the gender gap, it could be that we measured risk attitudes on the wrong lotteries, since perceived lotteries may be very different.

Basically, the problem is that choices of participants depend in a very intricate way on risk aversion, which may be hard to capture, As such, indirect approaches like we took them in our paper, may be more reliable, as they circumvent the issue of measurement, or making the right assumptions about how risk aversion enters the decision precisely (in combination with beliefs about performance, actual performance, etc.).

CONCLUSIONS

[19] This difference is marginally significant with a two-sided Fisher's exact test (p= 0.100).

[20] This difference is marginally significant with a two-sided Fisher's exact test (p= 0.092).

In this chapter, I wanted to show how to use experiments to test theory. I showed how the theory can be tested at deeper levels, by attacking the assumptions of the theory directly. I also provided two examples of design to test the theory: the two way design, and the elimination design. In the two way design, the initial environment, which allowed for two competing theories to explain the initial results, is changed in a way to separate the two theories. In the elimination design, the explanatory power of a theory for a phenomenon is questioned by changing the environment in a way that one the theory has no explanatory power anymore, and second, the phenomenon is still present. This at least casts doubt that the initial theory was the main driving factor for the result.

Finally, intelligent design can also be used when the initial hypothesis is not grounded in a careful model. In fact, most examples of intelligent design I presented in this chapter do not really deeply rely on the parameters of a model, but rather exploit broad results or assumptions. Often a more direct approach may help us to learn more.

REFERENCES

Andreoni, James and Lise Vesterlund, “Which is the Fair Sex? On Gender Differences in Altruism,” *Quarterly Journal of Economics*, 116, February 2001, 293-312.

Beyer, Sylvia, “Gender Differences in the Accuracy of Self-Evaluations of Performance,” *Journal of Personality and Social Psychology*, LIX (1990), 960–970.

Beyer, Sylvia, and Edward M. Bowden, “Gender Differences in Self-Perceptions: Convergent Evidence from Three Measures of Accuracy and Bias,” *Personality and Social Psychology Bulletin*, XXIII (1997), 157–172.

Bosch-Domènech, Antoni, Jose García-Montalvo, Rosemarie Nagel and Albert Satorra “One, Two, (Three), Infinity…: Newspaper and Lab Beauty-Contest Experiments”, *American Economic Review,* Dec. 2002, Vol. 92 (5), pp 1687-1701.

Byrnes, James P., David C. Miller, and William D. Schafer, “Gender Differences in Risk Taking: A Meta-Analysis,” *Psychological Bulletin*, LXXV (1999), 367–383.

Camerer, Colin, Teck Ho and Juin-Kuan Chong "A Cognitive Hierarchy Model of One-Shot Games," *Quarterly Journal of Economics* 119: 3 (August 2004), 861-898.

Camerer, C. and Lovallo, D. (1999). “Overconfidence and excess entry: an experimental approach’*, American Economic Review*, vol. 89(1) (March), pp. 306–18.

Casari, Marco, John C. Ham and John H. Kagel, “Selection Bias, Demographic Effects, and Ability Effects in Common Value Auction Experiments”, *American Economic Review*, September 2007, Vol 97 (4), 1278-1304.

Costa-Gomes, Miguel A. and Vincent P. Crawford, “Cognition and Behavior in Two-Person Guessing Games: An Experimental Study,” *American Economic Review* 96 (December 2006), 1737-1768;

Costa-Gomes, Miguel A. and Vincent P. Crawford and Nagore Iriberri, "Comparing Models of Strategic Thinking in Van Huyck, Battalio, and Beil’s Coordination Games" *Journal of the European Economic Association* 7 (2009), 377-387.

Costa-Gomes, Miguel A. and Georg Weizsaecker, "Stated beliefs and play in normal form games", 2008, *Review of Economic Studies* 75, 729-762.

Cox, James C., Vernon L. Smith and James M. Walker, “Theory and Individual Behavior of First-Price Auctions, *Journal of Risk and Uncertainty*, March 1988, 1, 61-99.

Cox, James C., Vernon L. Smith and James M. Walker, “Theory and Misbehavior of First-Price Auctions: Comment”, *American Economic Review*, Vol. 82, No. 5 (Dec., 1992), pp. 1392-1412

Crawford, Vincent P. and Nagore Iriberri, “Level-k Auctions: Can a Non-Equilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private –Value Auctions,” *Econometrica* 75, November 2007, 1721-1770.

Croson, Rachel, and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature*, 47(2): 448–74.

Eckel, Catherine C., and Philip J. Grossman, “Sex and Risk: Experimental Evidence,” in Handbook of Experimental Economics Results, C. Plott, and V. Smith, eds. (Amsterdam, The Netherlands: Elsevier Science B.V./North-Holland, forthcoming, 2002).

Erev, Ido, Eyal Ert, and Alvin E. Roth, “A choice prediction competition for market entry games: An introduction,” *Games*, Special Issue on Predicting Behavior in Games, 2010, 1 (2), 117-136.

Erev, Ido, Eyal Ert, Alvin E. Roth, Ernan Haruvy, Stefan Herzog, Robin Hau, Ralph Hertwig, Terrence Steward, Robert West, and Christian Lebiere, “A choice prediction competition, for choices from experience and from description,” *Journal of Behavioral Decision Making* , special issue on Decisions from Experience, 23, 1 (January) 2010 , 15 – 47.

Erev, I., Roth, A.E.: “Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria.” *American Economic Review* 88(4), 848–881 (1998)

Erev, Ido, Alvin E. Roth, Robert L. Slonim, and Greg Barron, "Learning and equilibrium as useful approximations: accuracy of prediction on randomly selected constant sum games," *Economic Theory*, special issue: Behavioral Game Theory Symposium, 33, October 2007, 29-51.

Eyster, Eric and Matthew Rabin, 2005, “Cursed Equilibrium”, *Econometrica*, 73(5), 1623-1672.

Friedman, Daniel, “Theory and Misbehavior of First-Price Auctions: Comment”, *American Economic Review*, Vol. 82, No. 5 (Dec., 1992), pp. 1374-1378

Georganas, Sotiris, Paul J. Healy and Roberto Weber, “On the Persistence of Strategic Sophistication”, working paper, 2009.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini, “Performance in Competitive Environments: Gender Differences,” *Quarterly Journal of Economics*, CXVIII, August 2003, 1049 – 1074.

Grosskopf, Brit, and Rosemarie Nagel, “The Two-Person Beauty Contest,” *Games and Economic Behavior* 62 (2008) 93–99.

Hertwig, Ralph and Andreas Ortmann. 2001. “Experimental Practices in Economics: A Methodological Challenge for Psychologists?” *Behavioral and Brain Sciences* 24:383–451.

Haile, Philip, Ali Hortaçsu and Grigory Kosenok, “On the Empirical Content of Quantal Response Equilibrium”, *American Economic Review*, March 2008, 98(1), 180-200.

Harrison, Glenn W., “Theory and Misbehavior of First-Price Auctions,” The American Economic Review, Vol. 79, No. 4 (Sep., 1989), pp. 749-762.

Harrison, Glenn W., “Theory and Misbehavior of First-Price Auctions: Reply,” *American Economic Review*, Vol. 82, No. 5 (Dec., 1992), pp. 1426-1443.

Ho, Teck-Hua, Colin Camerer, and Keith Weigelt, “Iterated Dominance and Iterated Best response in p-Beauty Contests,” *American Economic Review*, LXXXVIII (1998), 947–969.

Hoelzl, Erik and Aldo Rustichini, “Overconfident: Do You Put Your Money On It?,” *Economic Journal*, 04 2005, 115 (503), 305–318.

Ivanov, Asen, Dan Levin and Muriel Niederle, “Can Relaxation of Beliefs Rationalize the Winner’s Curse? An Experimental Study”, *Econometrica*, July 2010, Vol. 78, No 4, 1435-1452.

Jehiel, P. (2005), “Analogy-Based Expectation Equilibrium,” *Journal of Economic Theory*, 123, 81-104

Jehiel, P. and F. Koessler (2008), “Revisiting Games of Incomplete Information with Analogy-Based Expectations,” *Games and Economic Behavior*, 62, pp. 533-557.

Kagel, J. H. 1995. Auction: Survey of experimental research. In Alvin E. Roth and John H. Kagel, Editors, Princeton University Press.

Kagel, John H. and Dan Levin, “Independent Private Value Auctions: Bidder Behavior in First-, Second-, and Third-Price Auctions with Varying Numbers of Bidders”, *The Economic Journal*, Vol. 103, No. 419, (July 1993), 868-879.

Kagel, John H. and Dan Levin, “Auctions: A Survey of Experimental Research, 1995 – 2008,” In Alvin E. Roth and John H. Kagel, Editors, Princeton University Press, forthcoming.

Kagel, John H. and Alvin E. Roth, “Theory and Misbehavior in First-Price Auctions: Comment” *The American Economic Review*, Vol. 82, No. 5 (Dec., 1992), pp. 1379-1391.

Lichtenstein, Sarah, Baruch Fischhoff, and Lawrence Phillips, “Calibration and Probabilities: The State of the Art to 1980,” in Judgment under Uncertainty: Heuristics and Biases, Daniel Kahneman, Paul Slovic, and Amos Tversky, eds. (New York: Cambridge University Press, 1982).

McKelvey, Richard D. and Thomas R. Palfrey. 1995. “Quantal Response Equilibria for Normal Form Games.” *Games and Economic Behavior*, 10(1), 6-38.

Merlo, Antonio and Andrew Schotter, “Theory and Misbehavior of First-Price Auctions: Comment”, *American Economic Review*, Vol. 82, No. 5 (Dec., 1992), pp. 1413-1425.

Mobius, Markus M, Niederle, Muriel, Niehaus, Paul and Tanya Rosenblat, “Maintaining Self-Confidence: Theory and Experimental Evidence”, working paper 2010.

Nagel, Rosemarie, 1995. "Unraveling in Guessing Games: An Experimental Study," *American Economic Review*, vol. 85(5), December, pages 1313-26.

Niederle, Muriel and Alvin E. Roth, “Market Culture: How Rules Governing Exploding Offers Affect Market Performance," *American Economic Journal: Microeconomics*, 1, 2, August 2009, 199-219.

Niederle, Muriel, and Lise Vesterlund, “Do Women Shy Away from Competition? Do Men Compete too Much?,” *Quarterly Journal of Economics,* August 2007, Vol. 122, No. 3, 1067-1101.

Niederle, Muriel and Lise Vesterlund, “Explaining the Gender Gap in Math Test Scores: The Role of Competition,” *Journal of Economic Perspectives*, Spring 2010, Vol 24, Number 2, 129-144.

Roth, A.E., "Let’s Keep the Con out of Experimental Econ.: A Methodological Note" *Empirical Economics* (Special Issue on Experimental Economics), 1994, 19, 279-289.

Rubinstein, Ariel, “Dilemmas of An Economic Theorist,” *Econometrica*, 74 (2006), 865-883.

Stahl, Dale O., “Is Step-j Thinking an Arbitrary Modeling Restriction or a Fact of Human Nature?” *Journal of Economic Behavior and Organization*, XXXVII, (1998), 33–51.

Stahl, Dale and P. Wilson, 1995, “On Players’ Models of Other Players: Theory and Experimental Evidence”, *Games and Economic Behavior*, 10, 218-254.