Experimental economics

Lecture 4: Experimental design

Matej Lorko
matej.lorko@euba.sk
Materials: www.lorko.sk/lectures

References:

- Weimann, J., & Brosig-Koch, J. (2019). Methods in experimental economics. Springer International Publishing. Chicago
- Jacquemet, N., & l'Haridon, O. (2018). Experimental economics. Cambridge University Press.

Methodological foundations

- The times are long gone when economists “just did an experiment” to see what happens when you let subjects make specific decisions. In the meantime, methodological standards and procedures have evolved. Following these procedures is an important prerequisite for obtaining experimental results that can claim to meet the scientific standards of the economics profession.

- What is actually a good experiment? A good experiment succeeds in testing the most important effect a theory describes, while at the same time controlling for all plausible alternative explanations. This is based on the fact that there can often (if not always) be more than one potential explanation for an empirically observable regularity. Good experiments are undoubtedly those that, from the many alternative causal relationships, can find the one that actually explains the observed phenomenon.

- It should be borne in mind, however, that this is a very ambitious goal that can only relatively rarely be achieved. This is simply because not all the alternatives are always known. Therefore, the design of an experiment should eliminate the possibility that the results are influenced by factors that are not connected to the assumed causal relationships. If a faulty design or the inept execution of an experiment is responsible for the observed regularity, then one has certainly not carried out a good experiment.

- For this reason, it is worth considering the way in which various elements of an experiment’s design can influence the behavior of the subjects. This is the only way to be certain that the results obtained are an empirically significant regularity and not just an artifact of a bad experiment.

Experimental design

- If a research question is to be answered experimentally and with the aid of statistical methods, the experiment must be designed in such a way that it answers this question as well as possible. “As well as possible” in means that the choice of method has been made in such a way that the formal method of analysis is appropriate to the statistical nature of the data generated so that they are compatible.

- Before we send subjects to the laboratory to generate raw data for us and before we commit ourselves to some statistical method of analysis, the research question, design of the experiment, the resulting raw data and the statistical analysis must be precisely matched with each other.

- A poorly designed experiment leads to a weak scientific result – even the most sophisticated method of analysis cannot change that. On the other hand, a well-founded analytical method can derive an even more significant scientific insight from a well-designed experiment.

Experimental design

- From a statistical point of view, the course of an experimental study should be divided into a design phase and an execution phase. The design phase, which is to be carried out first, consists of the following tasks and typical issues:

- Operationalizing the research question: What are the central constructs for which data must be collected during the experiment in order to answer the research question? Can these constructs be measured as variables? How should these variables be measured? Which of them is the dependent variable? Which of them are independent variables?

- Structuring the statistical design:

  - Which variables are to be manipulated in which way by the experimenter (choice of treatments)?

  - Which variables can I control and how can an undesired variation of the dependent variable be minimized?

  - What is the observational unit and what is the experimental unit?

  - How should a sample of subjects be selected? How many subjects do I need to show correctly that a certain effect “exists” with a given probability? What groups of subjects should be formed and what method should be used to form these groups?

  - Will variables be measured on several levels (e.g. within-subject and between-subject)?

  - How frequently and when should each subject’s variable be measured? Which are the qualitative variables and which are the quantitative variables?

- Translating the research question into a statistical hypothesis or a statistical model: What formal relationship could exist between the observed variation of the dependent variable and the variation of the independent variables? Which are the fixed-effect variables and which are the random-effect variables?

Types of Variables

- In order to test a research idea experimentally, it is necessary to generate different types of variables. For example, suppose that the research hypothesis is that “the amounts offered in the ultimatum game are lower if the first mover is playing against a computer instead of a human being (and he or she knows this)”. In this case, the dependent variable is the amount offered by the first mover.

- An independent variable is expected by the experimenter to have an influence on the dependent variable, but not vice versa. In accordance with our research hypothesis, we expect the binary variable “computer opponent” (yes/no) to have an impact on the amounts offered. In a controlled experiment, the values of these independent variables are set systematically rather than simply being observed by the experimenter. In the above example, the experimenter measures the dependent variable “amounts offered” once under the value “yes” and once under the value “no” so that a comparison of both conditions is possible and the research hypothesis can be tested. In this case, the independent variable is also called a treatment variable, because its values represent the “treatments” or comparison conditions of the experiment under which the dependent variable is observed.

- Some further points need to be considered if the study is to draw a causal conclusion about the dependent and independent variables (and this is the main purpose of controlled experiments). If we observe a difference in the amounts offered once under each of the conditions “computer opponent – yes” and “computer opponent – no”, we must be able to rule out that this difference was caused by other influences - counfounding variables. Confounding variables blur the causality between dependent and independent variables because they have a “hidden” influence on the dependent variable that is not explicitly part of the experiment.

- Unfortunately, there are also confounding variables that cannot be controlled for. These are mainly such factors that make up the individual personality of a subject. Examples are intelligence quotient, income of parents, allergies, education, political sentiments, spatial ability, physical fitness and many more. Of course, not all possible uncontrollable variables are relevant to our own experiment, since many have no connection whatsoever to our dependent variable. Nonetheless, we would be well advised to carefully consider what, on the one hand, has a high probability of influencing our dependent variable and, on the other hand, can vary from subject to subject while at the same time remaining beyond our control.

Measurement

- Before testing hypotheses, we must understand some issues involving the measurement of the concepts we have decided to investigate and how we record systematic observations using numerals or scores to create variables that represent the concepts for analysis.

- How researchers measure their concepts can have a significant impact on their findings; differences in measurement can lead to totally different conclusions.

- It is useful to think of arriving at the definition of the variables as being the last stage in the process of defining a concept precisely. We often begin with an abstract concept (such as democracy), then attempt to define it in a meaningful way, and finally decide in specific terms how we are going to measure it.

- At the end of this process, we hope to attain a definition that is sensible, close to our meaning of the concept, and exact in what it tells us about how to go about measuring the concept.

- To be useful in providing scientific explanations for behavior, measurements of phenomena must correspond closely to the original meaning of a researcher's concepts.

- They must also provide the researcher with enough information to make valuable comparisons and contrasts. Hence, the quality of measurements is judged in regard to both their accuracy and their precision.

Accuracy of measurement

- There are two major threats to the accuracy of measurements. Measures may be inaccurate because they are unreliable and/or because they are invalid.

- Reliability describes the consistency of results from a procedure or measure in repeated tests or trials. In the context of measurement, a reliable measure is one that produces the same result each time the measure is used. An unreliable measure is one that produces inconsistent results-sometimes higher, sometimes lower.

- The reliability of social science measures can be calculated in many different ways.

  - The test-retest method involves applying the same "test" to the same observations after a period of time and then comparing the results of the different measurements.

  - The alternative-form method of measuring reliability also involves measuring the same attribute more than once, but it uses two different measures of the same concept rather than the same measure.

  - The split-halves method of measuring reliability involves applying two measures of the same concept at the same time. The results of the two measures are then compared. This method avoids the problem that the concept being measured may change between measures.

- A valid measure is one that measures what it is supposed to measure. Unlike reliability, which depends on whether repeated applications of the same or equivalent measures yield the same result, validity refers to the degree of correspondence between the measure and the concept it is thought to measure.

Control, Randomization and Sample Size

- Regardless of whether or not an uncontrolled confounding variable is measurable, its impact on our dependent variable should be removed from the experiment as far as possible; otherwise a clear causal conclusion with respect to our treatment variable is no longer possible. A 100% control of such variables is hardly possible since many of them are not only not measurable, but also unknown and their influence is therefore “hidden”.

- Nevertheless, there is a simple statistical trick that can mitigate their impact. The basic idea is to form two groups of subjects across which the possible confounding factors are distributed as evenly as possible. This is done by randomly assigning each subject to one of the groups (randomization). In the process, it should be ensured that the groups consist of a sufficiently large number of independent subjects.

- All in all, in a laboratory experiment, the central variable is the dependent variable. Changes in this variable are due to the influence of explanatory variables and various confounding factors. If the observed change in the dependent variable is to be attributed to a change in the explanatory variable induced by the experimenter, the three most important concepts to be considered are:

  - Control (all the unwanted influences that can be kept constant should be kept constant);

  - Randomization (create comparison groups that are homogeneous on average by leaving it to chance which subject is placed into which group);

  - Sample size (or replication) - ensure a sufficient number of independent observations in a treatment, i.e. sufficiently large groups of subjects who do not systematically exhibit the same behavior.

Creating the Statistical Design

- Compiling the Observation Units

- Selecting a certain number of subjects from a total population is referred to as sampling in statistics. Some thought needs to be given to the sample size, i.e. the question “How many subjects do I draw from the specified population?” Unfortunately, in experimental practice this question is often answered solely on the basis of the budget, true to the motto: “We simply take as many subjects as we can pay for, regardless of whether this number is large or small enough”. In the neurosciences, for instance, laboratory times are extremely expensive, so that sample sizes are (often have to be) in the single-digit range.

- However, such small samples are problematic, especially from the point of view of inferential statistics. The probability that a statistical hypothesis test correctly identifies an actual effect as present (this is called the power of a test) decreases drastically with smaller samples. In other words, even if in reality there is a relatively strong and scientifically relevant effect in the population, it will at best be recognizable as a “random artifact” and not as a statistically significant effect. On the other hand, there is also a “too large” in terms of sample size, since having samples that are too large can make statistical hypothesis tests too sensitive. This means that even the smallest, possibly scientifically insignificant effects become statistically significant.

- It is thus already clear that statistical significance should not be confused with scientific significance. Depending on the sample size, both can be completely different. This is because statistical significance is strongly influenced by the sample size, whereas the true effect to be detected in a population is not.

- If it is clear that a (sufficiently large) random sample is not affordable and a representative sample is still required, then stratified sampling is a good possibility. The population is first divided into subpopulations (strata), with the subjects within each subpopulation having at least one common characteristic that distinguishes them from the subjects of the other subpopulations. A random sample is then drawn from each stratum. Each of these samples must make up the same proportion of the total of all samples as each stratum in the total population.

Sampling

- Suppose we want to assess national level of support for some proposed government policy. Since it is impossible to interview everyone, a more practical approach is to select just a "few" members of the population for further investigation. This is where sampling comes in.

- A sample is any subset of units collected in some manner from a population. The sample size and how its members are chosen determine the quality (that is, the accuracy and reliability) of inferences about the whole population.

- A researcher's decision whether to collect data for a population or for a sample is usually made on practical grounds. The advantages of taking a sample are often savings in time and money. The disadvantage is that information based on a sample is usually less accurate or more subject to error than is information collected from a population.

- Once a sample has been gathered, features or characteristics of interest can be examined and measured. The attributes of most interest in empirical research are numerical or quantitative indicators such as percentages or averages. These measures, or sample statistics, as they are known - are used to approximate the corresponding population values, or parameters.

- In order to mitigate the sample bias, ideally each element in the total population should have a known probability of being included in the sample. This knowledge allows a researcher to calculate how accurately the sample reflects the population from which it is drawn.

What can be learned from samples

- Samples provide only estimates or approximations of population attributes. Occasionally these estimates may be exactly right, but most of the time, however, they will differ from the true value of the population parameter.

- When we report a sample statistic, we always assume there will be a margin of error, or a difference between the reported and actual values.

- Where does the loss of precision or accuracy come from? The answer is chance, or luck of the draw. If you flip a coin ten times, you probably won't get exactly five heads, even if the coin is fair or the probability of heads is one-half. Randomness seems to be an innate feature of nature, at least on the scale at which we observe it.

- Just as with our coin toss, a random sample of ten (or even much larger) is not likely to produce precisely the value of a corresponding population parameter. But if we follow proper procedures and certain assumptions have been met (for example, the sample is a simple random sample from an infinite population), a sample statistic approximates the numerical value of a population parameter.

Creating the Statistical Design

- How Do Experimental Treatments Differ?

- It is possible to classify experimental treatments according to the number of factor variables and their type as well as the number of possible values. In a single factorial design, only a single variable is changed. If this is a binary variable with just two values, or levels, we speak of a 1×2 factorial design. 1×2 factorial designs can be evaluated particularly easily since only the mean values of the dependent variables are usually compared under these two treatment conditions. Ideally, this difference is due to the treatment itself and is therefore called the (simple) treatment effect. The quantitative difference between the two values is called the size of the treatment effect or the (unstandardized) effect size. If, on the other hand, the factor variable has more than two levels, the treatment is called multilevel factorial design. In this case, the mean values of the dependent variable can be compared pairwise for every two levels or simultaneously for all levels.

- A design with two factors is considerably more complex than a single factorial design. For example, if we want to experimentally investigate how the factors “games against the computer” (Comp: no/ yes or 0/1) and “the experimenter knows who I am” (Anon: no/yes or 0/1) affect the giving behavior in a dictator game, then this hypothetical 2×2 factorial design.

- In the repeated measures design, each subject undergoes several measurements, either in one and the same treatment at different times (longitudinal design) or in different treatments, naturally also at different times (cross-over design). The sequence of treatments a subject goes through is again randomized. In each case, multiple measurements generate a within-subject structure with several observations for each subject.

- The main statistical problem with multiple measurements is the interdependence of the observations. In a 1×2 factorial design with multiple measurements, we get a control group (measured at level 1) and a treatment group (measured at level 2), which are related. Thus, the effect measured using the dependent variable can no longer be clearly attributed to the treatment, since it could just as easily be a time or sequence effect (e.g. learning, familiarization, fatigue). Counterbalancing the order (balancing) often comes to our aid in this case, i.e. two homogeneous groups are formed and one group is measured in the order level 1, then level 2 and the other in the order level 2, then level 1.

- The advantages of repeated measurements are lower costs due to fewer subjects, lower error spread, thus resulting in higher statistical power than comparable between-subject designs, and the possibility of measuring treatments over time (dynamics). The disadvantages of such a design are that it involves considerably more complex methods of analysis due to the dependency of the observations and weaker causalities owing to sequence, time and carry-over effects.

Within- Versus Between-Subject Design

- At the core of experimental research stands the comparison of different experimental treatments under controlled conditions. An experiment that consists of only one treat- ment makes relatively little sense. It is almost always a case of subjects making decisions under different conditions, with the treatments that are being compared as far as pos- sible differing in only one parameter, thus enabling conclusions in relation to causality to be made.

- A fundamental issue of experimental design in this regard is whether each individual subject participates in a number of different treatments or whether every treatment involves different subjects, with each subject participating in only one treat- ment. The first case is described as a “within-subject design”, since the comparison takes place within one and the same subject, while the latter case is called “between-subject design” due to the comparison between the subjects.

- Within-Subject design

- Advantages: the number of observations per subject is greater when each subject participates in several treatments than when new subjects are invited, the internal validity of the experiment does not require successful randomization to have been carried out, closer proxmity to theory - higher external validity.

- Disadvantages: there is no avoiding that dependencies arise between the individual observations in the different treatments (can be solved by taking into acccounts the order effects, and by panel data analysis), presenting the subjects with different treatments can lead to an experimenter demand effect

- Between-Subject design

- Advantages: easy to handle - the only condition to be met is that the subjects are randomly assigned to the different treatment. the statistical analysis of between data is easier than within data, as it is not necessary to correct for dependencies between data elements. Also, between designs tend to lead to conservative results as compared to within designs, therefore there is a relatively high certainty that this finding is revealing a causality.

- Disadvantages: considerably more resources (time, money, subjects) may be required to obtain statistically meaningful data than with the within design approach. In other words, with the same use of resources, less statistical “power” is likely to be achieved with a between design than with a within design. Moreover, the external validity is not as direct as with a within design.

Strategy Method Versus Direct Response

- It is generally easy to determine the elicitation method to use in experiments involv- ing the decision-making behavior of individual subjects without the occurrence of any strategic interaction. The subjects are presented with a specific decision problem, i.e. they have to make a choice, and it is this choice that is observed. The matter can become much more complex if strategic interactions arise in the experiment. It is, in the first instance, irrelevant whether the game played by the subjects takes place simultaneously or sequentially. For better understanding, however, it is simpler to assume a sequential game.

- Direct elicitation (“hot”) vs. with the strategy method (“cold”).

- The normal case is that the players make their moves in the order specified, with the second mover responding to the move made by the first mover, the third mover react- ing to that of the second mover, etc. The players thus provide a direct response to the action of the mover before. This method of eliciting the responses is simple and easy to understand. From the point of view of the experimenter, however, it can have a considerable drawback.

- Let us take the simplest sequential game imaginable. Two players each choose between two possible alternatives. In this case, there are four possible outcomes of the game. Each individual decision that is observed, however, only provides information about one of the four possible paths on which the game tree can be traversed. Suppose the first mover has a choice between alternatives a and b. If the first mover (for whatever reason) has a preference for a, and chooses this strategy in nine out of ten cases, it becomes quite difficult and expensive to collect enough observations in the subgame following b.

- The strategy method, which essentially goes back to an article by Selten (1967), offers an elegant solution to this problem. Instead of the second mover being presented with the decision of the first mover, he is required to specify a complete strategy. In our simple example, he has to indicate what he will do at the two decision-nodes he can reach. In other words, he must indicate how he will respond in both cases, i.e. if first mover plays a and if he plays b. The result of the game is obtained by combining the move chosen by the first mover with the corresponding response from the strategy of the second mover. In this way, the experimenter elicits information about behavior throughout the game.

Actions and rewards

- At the heart of all experimental investigations is the ability to make observations under controlled conditions. For example, experiments with animals attempt to study their learning ability. For this purpose, certain signals are combined with rewards (in the form of food) and the researchers observe whether the laboratory animal is able to make a connection between the signal and the reward associated with it.

- By controlling the variation of signals and rewards, conclusions can be drawn about the animals' ability to learn. It is implicitly assumed in such experiments that the animals like the food offered as a reward and that they prefer to have more of it. Only then is it justifiable to assume that the animals are making an effort (learning is strenuous) to obtain the food. Now the assumption that apes, for example, like to eat sweet fruit and are particularly keen on certain "treats" is not too daring. It is easy to see that this is the case. The behavioral hypothesis for an ape is therefore "I prefer more bananas to fewer bananas".

- But what about experiments with humans? Ultimately, economic theory describes above all else how people make decisions. Of central importance are the preferences that are attributed to the actors. Rational choices are always related to the goal being pursued. Thus, it is not possible to make a prediction about what a rational decision-maker will do if it is not known which goal this person is pursuing.

- Experiments that test theories and also experiments that are not based on theory would therefore be worthless without assumptions about the underlying preferences. The problem is that people's preferences are doubtlessly more differentiated than those of apes. In other words, it makes little sense to assume that people prefer more bananas to fewer bananas. Fortunately, however, the difference between us (humans) and our closest genetic relatives is not so great, thanks to the fact that there is a banana equivalent of sorts for us, and that is money.

Induced value method

- Vernon Smith introduced this equivalent systematically into experimental methodology in 1976 and gave it a name: the induced value method. The idea is very simple. It is assumed that the consumption of every good generates utility, for which there is a monetary equivalent – the willingness to pay for the good in question. If each utility value can be expressed in terms of money, then the utility function can also be replaced by a "money function", and by introducing this money function into the experiment as a "payoff function", one has induced from the outside the utility function that is used for the evaluation of options for action. The following example best illustrates this process.

- The induced value method requires that people react to money in the same way as apes react to bananas – more of which is always better than less. This is also one of a total of three requirements that Smith specifies need to be met for the induced value method to be applied. First, the utility function must grow monotonically in terms of money. In slightly more technical terms, if a decision-maker can choose from two alternatives and one of them has a higher payoff than the other, then the decision-maker will always choose the alternative with the higher payoff.

- Second, the payoffs have to be salient. The so-called salience requirement is understood as meaning that the decision to be taken by a subject in an experiment must also be payoff-relevant. It is worth considering a little more deeply what this implies. It is important that the payoff function is not too flat. If it is, taking different decisions has little impact on the resulting payoffs, which can result in subjects not putting much effort into actually making what for them is the best decision, because a mistake has little financial impact.

- The third requirement that Smith lists is the dominance of the payoffs. As an experimenter, one has to be aware of the fact that experimental subjects could also have other things on their minds than the money they can earn in an experiment. For example, people dislike getting bored or thinking that their precious time is being wasted. Boredom may lead to subjects having an incentive to make things more interesting by trying things out without paying too much attention to their payoffs. There are many other factors that might discourage subjects from focusing exclusively on maximizing their payoffs.

- They could form expectations about what the experimenter wants from them and behave accordingly (the experimenter demand effect mentioned earlier). They might also make social comparisons and try to outperform the other subjects by doing things that adversely affect them. They may even develop altruistic feelings or think about fairness. All this and much more is possible. What Smith means by "dominance" is that, despite all these distractions, the pursuit of the highest possible payoff still comes first, and in case of doubt, the alternative that ensures the highest payoff is chosen.

The Size of Payoffs

- It's about money. But how much money we are actually talking about is a question we have sidestepped a little so far. Do monetary incentives have an effect and if so, to what extent does it depend on the size of the payoffs? This is a question that is frequently directed at experimental economic research. Two extreme positions are conceivable. One of them assumes that it makes no difference at all whether monetary incentives are used or not. The other extreme position is that all the deviations from the model of rational choice that can be observed in experiments disappear if the payoffs are set at a sufficiently high level.

- Camerer and Hogarth (1999) analysed 74 papers in which the effect of different payoff levels was investigated. The most important message from their study is that the two extreme positions described above are wrong. Monetary incentives are not ineffective (i.e. decisions should not be elicited hypothetically, but provided with appropriate mon- etary consequences) and deviations from the rational choice model do not disappear at higher payoffs. The last point can be expressed a little more precisely. The authors have not found a single study where a deviation from the rational choice model observed at low payoffs disappears when the payoff is increased.

- Nevertheless, the work of Camerer and Hogarth also shows that the effect of incentives is not always the same. It may well depend on the special circumstances of the experiment. For instance, an increase in payoffs has an impact if the payoff a subject receives at the end of the experiment depends on the effort involved. A good example is provided by experiments testing memory ability. Here it is profitable for the subjects to be more attentive and the more they can earn, the more attentive they actually are.

- The size of the payoff, on the other hand, has no influence in experiments in which the subjects already have a sufficiently high level of intrinsic motivation or in which any additional effort is not worthwhile because the payoff function is flat. Such experiments show, however, that the variance decreases, i.e. the average amount given remains the same, but there are fewer deviations upwards or downwards. The size of the payoff also has little influence on the behavior under risk. At best, there is a slight tendency towards more risk-averse behavior.

- The following rule of thumb should suffice: it is necessary to set noticeable but not exorbitantly high incentives. As a rule, the size of the payoff should be based on the opportunity costs of the subjects in the experiment.

Experiments with Real Effort

- Economic experiments almost always involve decisions in which costs play a role, whether it is a case of the subjects being faced with an allocation task in which every amount they give is at the expense of their payoff, purchasing goods or making a contribution to the production of goods. Occasionally, the work efforts that are exerted to fulfill a task are also represented by appropriately designed cost functions (for example, in the minimum effort coordination game). A two-stage procedure is usually used to implement costs in the laboratory. The first stage consists of giving the subjects an income in the form of an initial endowment (house money). This income can then be used to cover the costs incurred. In the second stage, the costs are specified in the form of a mathematical function, with there being considerable room for creativity. For example, the cost function can be convex to represent that it becomes increasingly difficult to exert the effort.

- Inducing costs in this way has considerable advantages, especially in view of the fact that the experimenter retains complete control. Since the costs are part of the payoff function, it is indisputable to what extent they are actually incurred. However, this high degree of control comes at a price. People may treat the money they are given differently from the money they have earned from work. It is therefore not entirely unproblematic to first give subjects money that they can then use to cover costs.

- An alternative to issuing house money is to have the subjects work for the money they receive by introducing real effort. This increases external validity and avoids the house money effect, but has the disadvantage that the control over costs is lost. If subjects are allowed to “work” in order to impose costs on them, the actual level of costs that the subjects incur depends on the burden of the work they have to bear – and that cannot be observed! The question is, under which conditions a real effort design is appropriate and, above all, how it can be designed in concrete terms.

- An important requirement is that it be structured in such a way that it can be assumed that at least at the beginning of the experiment all the subjects are equally good at achieving this performance. Therefore, no prior knowledge that may exist to varying degrees should be required and personal aptitude should not play an important role. It is also clear that the task should be easy to explain so that the subjects understand what is involved. Furthermore, the work outcome should be easily and reliably measurable and allow a comparison between the subjects. Finally, the task should be designed in such a way that possible learning effects are minimized and quantifiable, so that these effects can be corrected if necessary.

Selecting the Payoff Mechanism

- As a rule, economic experiments use monetary payoffs to create incentives in the labo- ratory, which are assumed either to be effective in the model (to be tested) or to play a role in real decisions. This raises not only the question of how large the incentives should be, but also how they should be paid. This question becomes much more important when the subjects make several decisions. In the last section, we dealt with experiments that deal with this very issue. In order that information about the risk preference of the subjects can be obtained, they are usually required to perform several lottery comparisons. However, repeated deci- sions or several similar decisions are not an exclusive feature of experiments to reveal risk preferences. On the contrary, they can be found in many contexts.

- At first glance, one might think that in such cases it is the gold standard to pay off all the decisions of all the subjects. Whether this standard is achieved solely depends on the funds available. But this point of view is wrong because the “pay-each-task” payoff method is only acceptable if it is ensured that the subjects of this method treat each indi- vidual decision as if they only had to make that one decision, thus making it necessary to examine each decision in isolation. However, there are good reasons for believing that in many cases this just cannot be guaranteed. Two effects can prevent this isolation hypothesis from being fulfilled.

- First, income effects can lead to decisions later in the experiment taking place under conditions that differ from those prevailing at the time of earlier decisions. If every decision is paid off individually, a subject can calculate how much he or she has already earned.

- The second effect that is capable of violating isolation is the portfolio effect. This means that in the case of decision under risk, the combined effect of decisions can lead to different results than if all individual decisions are taken separately. Take as an example the two-stage choice between two lotteries A and B, with the former being less risky than the latter. A risk-averse decision-maker would choose (A, A) for isolated decisions while a risk-seeker would choose (B, B). However, if the decision-maker can form a portfolio of both lotteries, it is possible that (A, B) has a higher expected utility than (A, A) and the risk-averse decision-maker therefore prefers (A, B)

- Wealth effects and portfolio effects can occur in repeated decisions in many cases and should therefore be eliminated by appropriately choosing the payoff mechanism. As a result, a lot of experiments pay for only one randomly selected decision, just as in Holt-Laury example.

Is It Okay to Take Money from Subjects of Experiments?

- In economic contexts, but also in other important situations in society, it is possible that decisions taken by people result in losses. Sometimes it is even the case that people may only be able to exert an influence on how high a loss is and are no longer able to avoid it altogether.

- An important and interesting question is whether decision-making behavior in the event of losses mirrors that of gains or whether there are systematic differences. The only way forward experi- mentally is to conduct experiments in which subjects actually face the risk of loss or even have to accept a loss with certainty. In such a case, the experimenter takes money away from the subject. Is he allowed to do this? Should he do this?

- It is sometimes claimed that it is unethical to take money from subjects of experiments if they make losses in the laboratory. This poses a dilemma for the experimenters. On the one hand, it is important to find out how people react to possible losses. On the other hand, experiments must be designed in such a way that the subjects end up receiving money and not having to pay anything. A popular method of overcoming this dilemma is to design experiments in such a way that, although there is a possibility that losses may occur in individual parts of the experiment, on average there will be no loss at the end of the experiment.

- Incorporating losses without any actual losses being incurred is one way of avoiding the dilemma. However, such approaches are compromises since they only reflect real losses in a limited sense. Not paying for all the losses that result in an experiment can bias decisions. A frequently implemented alternative to this is to pay the subjects a sufficiently high “show-up fee”, from which the potential losses can be paid.

The House Money Effect

- Monetary incentives are usually created in the experiment by, in a sense, pressing money into the hands of the subjects, who then can use it in the experiment. The basic idea here is that the value of money does not depend on where it comes from. Whether you work hard for 10 euros, find it on the street or win it in a lottery, it makes no difference to the quantity of goods you can buy for that money.

- So why should 10 euros received as a gift be worth less than the 10 euros earned? This view stems from rational choice theory. The notion that money always has the same value cannot be shaken, and under the requirements of the neoclassical rational choice model it would simply not be reasonable to value endowed money any differently from earned money.

- However, it obviously makes a difference whether the money used in an experiment is “your own money” or money provided by the experimenter. This endowed money is something like a windfall profit, i.e. income that simply lands in your pocket without you having to do anything about it.

- It is quite obvious that the unexpected gain changes your behavior – even if this is difficult to reconcile with rational behavior. It can be assumed that after a windfall profit, the propensity to consume increases just as much as the willingness to take risks. If that is the case, then of course this is highly relevant for the design of the payoffs in an experiment. The only question is how to experimentally test the effect of monetary gifts. How do we get people to use their own money in an experiment? Such an experiment is likely to make recruiting subjects quite difficult. For this reason, a different approach is taken.

- Instead of the subjects having to spend their own money, the experimenter has them perform a task for the money they receive. The type of task can be freely chosen. The crucial point is that the subjects no longer have the feeling that they have been given the money to use in the experiment. This is not quite the same as using their own, self-earned money, but if it turns out that in this sense money that is not endowed is treated differently from money that is, then it is safe to assume that it is the “house money effect” that describes this.

Is It Permissible to Lie to Subjects of Experiments?

- Experiments in which the subjects can suffer losses are very rare and even rarer are experiments in which they actually have to pay something. Most seldom, however, are experi- ments in which subjects are lied to – at least in experimental economic research. At first sight, this seems to be self-evident, since lies seem to be at least as unethical as asking for money. In light of this, it should be clear that such a thing is simply not done.

- On closer inspection, however, it can be seen that honesty in the laboratory is a specialty of economists and that there are other disciplines which are far from being as strict about this as economics. It is therefore worth taking a look at why economists insist on honesty and why, for example, experimental psychologists often fail to do so.

- There is a very broad consensus within the scientific community of experimental economists that deception cannot be tolerated. According to the argument, lying would lead to the experimenters gaining a reputation of not being honest. This, in turn, would have disastrous consequences because, if the subjects were to suspect that they were being lied to in the laboratory, how would it possible to monitor their preferences?

- If the experimenter did not know which game the subjects thought they were actually playing, he could, strictly speaking, no longer draw any conclusions from their behavior. Such a scenario must be prevented and can only be achieved by the experimenters defending their reputation of being honest.

- An important question does remain to be resolved in this connection, however. When does dishonesty begin? The rule could be formulated as follows. Everything that is said to the subjects must be true. However, the whole truth does not always have to be told all at once.

Are Students the Right Subjects?

- Well over 90% of all laboratory experiments are carried out with student subjects. And this not only applies to laboratory experiments in economics. Is this the right choice? This is a question which is constantly being raised with concern. Is it really possible to learn something about the behavior of people in general from the behavior of students? Or are students actually too “special”, i.e. not sufficiently representative?

- Students simply have many advantages. First of all, they are readily available, being represented in large numbers at universities and blessed with a relatively large amount of freely available time. That is why they can take part in an experiment, for example, at 2 in the afternoon or at 10 in the morning. Another advantage is that it can be assumed that students generally understand relatively easily and quickly what is expected of them in the experiment. From the point of view of the experimenters, it is also an advantage that students are often short of money and therefore gladly take the opportunity to earn something by participating in an experiment. The relatively low opportunity costs in terms of time mean that the monetary incentives set in the experiment do indeed carry a high weighting.

- These advantages are to a degree mirror images of the disadvantages of conducting experiments with non-student subjects. Recruiting the latter is much more difficult and time-consuming. If they are working people, only laboratory hours after work can be considered. In other words, it is necessary to get people to spend their scarce free time in the evening in the laboratory instead of at home with their families. In addition, it is difficult to establish initial contact. Students can be recruited relatively easily in lectures and are therefore usually represented in large numbers in a database of experimental subjects. Recruitment can be done at the push of a button or with a few clicks.

- This is not the case for non-students. Moreover, the opportunity costs are significantly higher for employed people than for students, with experiments with non-students, therefore, always being more expensive than those with students.

Are Students the Right Subjects?

- However, inviting students into your laboratory may create a double selection bias. First, students differ systematically from the average population. On the one hand, they are younger and better educated; on the other hand, they do not have the experiences of an average adult. For instance, they have no professional experience, do not know what it is like to pay income tax or to negotiate for their salary. These systematic differences make it difficult to transfer the decisions observed by students to the average population.

- The second selection bias comes into play when students participate voluntarily in experiments. It cannot be ruled out that only certain types of students participate in experiments. What is particularly worrying is that this self-selection process affects the preferences of the subjects, both their risk preferences and their social, or other regarding, preferences.

- Are students different from non-students? Exadaktylos et al. (2013) conclude that there is no significant difference in social behavior between student volunteers and non-student volunteers. These findings admittedly contrast with a whole series of observations showing significant differences between students and non-students.

- For example, Falk et al. (2013) found that in a trust game, non-students paid back significantly higher amounts to the first movers than students did. In the experiment already mentioned by Anderson et al. (2013), it can also be seen that students behaved much more selfishly than the adult volunteers and the truck drivers. Cappelen et al. (2015) examine social behavior in a dictator game experiment and in a trust game experiment. They also find significantly more pronounced social preferences in a group of subjects consisting of representative persons (of Norwegian society) compared to a group of students. Belot et al. (2015) come to the same conclusion. They also find that students are more likely to be able to think strategically than “normal citizens”.

- However, all in all, students are not the worst possible choice. The differences to the rest of the population tend to be moderate, while the use of experts in experiments is not unproblematic. Therefore, using students as subjects may well represent a good alternative in the vast majority of cases. This does not rule out the possibility that there may be specific questions in which it seems advisable to conduct experiments with a more representative population.

Survey Research and Interviewing

- How to ensure validity and reliability of survey and interview data?
- Let R stand for the respondent and I for the interviewer:
  - The requested information must be available to R (that is, not forgotten or misunderstood).
  - R must know what is to I a relevant and appropriate response.
  - R must be motivated to provide I with the information.
  - R must know how to provide the information.
  - I must accurately record R's responses.
  - The responses must reflect R’s meanings and intentions, not I's.
  - Other users of the data must understand the questions and answers the same way R and I do.

Survey research

- A group of individuals respond to or fill out more or less standardized questionnaires. The questionnaires may take different forms to investigate different hypotheses, but they do not involve freewheeling or spontaneous conversations.

- The validity of measures is higher when they are not clouded by other influences. If you get the chance to formulate the questions for a questionnaire, it is best to use questions from existing measures that are validated in previous research. Regardless of the origin of the items: make sure that they are formulated in short sentences that everyone will understand in the same way, and that the response options are exhaustive and easy to choose from. It will be good to use your own experience as a yardstick. Imagine that you are a respondent for the survey, would you be able to easily answer the questions? Would your responses to the items indeed represent your views?

- Although surveys can be relatively quick and cheap mean to obtain data, the researcher needs to think carefully about:

  - Completion rates - If the response rate is low, either because individuals cannot be reached or because they refuse to participate, the researchers' ability to make statistical inferences for the population being studied may be limited. Also, those who do participate may differ systematically from those who do not, creating other biases. Increasing the size of the survey sample to compensate for low response rates may only increase costs without alleviating the problem.

  - Sample-population congruence - how well the sample subjects represent the population, is always a major concern. Here we are speaking of how well the individuals in a sample represent the population from which they are presumably drawn. Bias can enter either through the initial selection of respondents or through incomplete responses of those who agree to take part in the study.

  - Questionnaire length - if a survey poses an inordinate number of questions or takes up too much of the respondents' time, the respondents may lose interest or start answering without much thought or care.

Response quality

- Response quality = the extent to which responses provide accurate and complete information. It is the key to making valid inferences. Response quality depends on several factors, including the respondents' motivations, their ability to understand and follow directions, their relationship with the interviewer and sponsoring organization, and, most important, the quality of the questions being asked.

- Engaging respondents - it is important to get off on a good footing by introducing yourself, your organization, your purpose, your appreciation of their time and .trouble, your nonpartisanship, your awareness of the importance of anonymity, and your willingness to share your findings.

- Since the whole point of survey research is to accurately measure people's attitudes, beliefs, and behavior by asking them questions, we need to spend time discussing good and bad questions.

- Good questions prompt accurate answers; bad questions provide inappropriate stimuli and result in unreliable or inaccurate responses. When writing questions, researchers should use objective and clear wording. Failure to do so may result in incomplete questionnaires and meaningless data for the researcher. The basic rule is this: the target subjects must be able to understand and in principle have access to the requested information.

- Certain types of questions make it difficult for respondents to provide reliable, accurate responses. These include double-barreled, ambiguous, and leading questions.

- Common problems in survey research are that respondents fail to report accurately about their behavior and give more positive answers to hypothetical questions than to questions about facts. Observational measures of behavior are to be preferred over self-reported measures. Factual questions (e.g., ‘Did you give to a charitable cause in the past month?’) are to be preferred over hypothetical questions (‘Would you give to a charity if asked?’). The problem that respondents report socially desirable attitudes and behaviors has often been studied as a ‘response bias’ (Meehl & Hathaway, 1946; Crowne & Marlowe, 1960).

Closed vs open-ended questions

- The main advantage of a closed-ended question is that it is easy to answer and takes little time. Another advantage is that answers are easy to compare, since all responses fall into a fixed number of predetermined categories. These advantages aid in the quick statistical analysis of data.

- With open-ended questions, by contrast, the researcher must read each answer, decide which answers are equivalent, decide how many categories or different types of answers to code, and assign codes before the data can be analyzed.

- Another advantage of closed-ended questions over open-ended ones is that respondents are usually willing to respond on personal or sensitive topics (for example, income, age, frequency of sexual activity, or political views) by choosing a category rather than stating the actual answer.

- Critics of closed-ended questions charge that they force a respondent to choose an answer category that may not accurately represent his or her position. Therefore, the response has less meaning and is less useful to the researcher.

- Also, closed-ended questions often are phrased so that a respondent must choose between two alternatives or state which one is preferred. This may result in an oversimplified and distorted picture of public opinion. A closed-ended question allowing respondents to pick more than one response (for example, with instructions to choose all responses that apply) may be more appropriate in some situations.