Journal of Economic Behavior & Organization 81 (2012) 1–8

Contents lists available at SciVerse ScienceDirect

Journal of Economic Behavior & Organization

journal homepage: www.elsevier.com/locate/jebo

Experimental methods: Between-subject and within-subject design

Gary Charness[a,*], Uri Gneezy[b], Michael A. Kuhn[c]

[a] Dept. of Economics, University of California at Santa Barbara, United States
[b] Rady School of Management, University of California at San Diego, United States
[c] Department of Economics, University of California at San Diego, United States

ARTICLE INFO

Article history:
Received 26 March 2011
Received in revised form 23 August 2011
Accepted 23 August 2011
Available online 21 September 2011

JEL classification:
B49
C91
C92

Keywords:
Within-subject
Between-subject
Experimental design and methodology

ABSTRACT

In this article we explore the issues that surround within-subject and between-subject designs. We describe experiments in economics and in psychology that make comparisons using either of these designs (or both) that sometimes yield the same results and sometimes do not. The overall goal is to establish a framework for understanding which critical questions need to be asked about such experimental studies, what authors of such studies can do to ameliorate fears of confoundedness, and which scenarios are particularly susceptible to divergent results from the two approaches. Overall, we find that both designs have their merits, and the choice of designs should be carefully considered in the context of the question being studied and in terms of the practical implementation of the research study.

1. Introduction

A fundamental characteristic of experimental approaches to economic studies is that researchers can observe behavior in an abstract environment that they control. Ideally, by exposing participants to different treatments, one is able to achieve identification of causality.

There are two primary ways in which experimenters can construct these environments. In a "within-subject" designed experiment, each individual is exposed to more than one of the treatments being tested, whether it be playing a game with two different parameter values, being treated and untreated, answering multiple questions, or performing tasks under more than one external stimulus. With such designs, as long as there is independence of the multiple exposures, causal estimates can be obtained by examining how individual behavior changed when the circumstances of the experiment changed. In a "between-subject" designed experiment, each individual is exposed to only one treatment. With these types of designs, as long as group assignment is random, causal estimates are obtained by comparing the behavior of those in one experimental condition with the behavior of those in another.

In this article we explore the issues that surround each of the two design approaches. We describe experiments in economics and in psychology that make comparisons using either within or between designs (or both) that sometimes yield the same results and sometimes do not. The overall goal is to establish a framework for understanding which critical questions need to be asked about such experimental studies, what authors of such studies can do to ameliorate fears of confoundedness, and which scenarios are particularly susceptible to divergent results from the two approaches.

* Corresponding author.
E-mail addresses: charness@econ.ucsb.edu (G. Charness), ugneezy@ucsd.edu (U. Gneezy).

0167-2681/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jebo.2011.08.009

Overall, we find that both designs have their merits, and the choice of designs should be carefully considered in the context of the question being studied and in terms of the practical implementation of the research study. In our view, between-subjects designs are more conservative and one should be cautious about carry-over and demand effects in within subjects designs; however, within designs lend themselves to more powerful econometric techniques and, in many cases, are a closer match to a theoretical perspective. We discuss how one might ameliorate the issues of concern regarding within designs.

In the remainder of this article, we provide an overview in Section 2 and some simple examples in Section 3. We discuss experiments where the two different methods led to different results in Section 4, and to similar results in Section 5. We describe some econometric issues in Section 6, and conclude in Section 7.

2. Overview

Both within and between designs (we will henceforth use "between" and "within") have their proponents. And yet it seems clear there are advantages and disadvantages to each approach, so the issue is more nuanced. Within designs may lead to spurious effects, through respondents expecting to act in accord with some pattern, or attempting to provide answers to satisfy their perceptions of the experimenter's expectations. This is known as a "demand effect"—according to which participants in experiments interpret the experimenter's intentions and change their behavior accordingly, either consciously or not (Rosenthal, 1976; White, 1977). Demand effects are likely to be stronger in a within design.

Within analyses have three main advantages with respect to between analyses. First, their internal validity does not depend on random assignment. Second, in many frameworks they offer a substantial boost in statistical power. Finally, they are more naturally aligned with most theoretical mindsets; a theorist is likely to imagine an agent in a market reacting to a price change, not two agents in separate markets with different prices. However, in environments where an individual is likely to only face a single decision, a between design might have more external validity. The disadvantages to within analyses are essentially a slew of confounds to identification that may be introduced because of the necessity of exposing each subject to multiple treatments. One has to worry about the order of exposure affecting the reference and framing of treatments.

We emphasize that this is a gross simplification of the distinction between the two approaches and that what constitutes exposure is critically dependent on the research question. For example, if the researcher's question is how exposure to X affects reactions to price changes, a between-subjects design involves observing two group of individuals react to price changes: one group in the presence of X and the other group outside of it. In such a between design, the fears that we have highlighted so far in the context of within designs are now present in both treatments. If exposure to X affects the formation of biases within subjects, then the between difference in behavior will not be the causal effect of X.

Between designs typically have no natural anchor. Thus, results inherently have substantial noise, and may miss important and real patterns. Real-world problems about whether to make a particular decision are often posed as between subjects; choices about which decision to make may be considered to be within subjects. Between analyses are statistically simple to perform as long as random assignment is achieved across groups. Little sophistication is required even when the games are extended beyond one round; if two groups play 20 rounds, and one group is treated while the other is not, we can compare between the two groups. The problem here is that statistical power is hard to come by because, in a strict sense, each group can only provide one independent data point. This is exacerbated by the fact that the nature of laboratory and field experiments generally lends itself to considerably smaller samples than is typically available with field-observational data and that between analyses have severe limitations in relation to testing a large parameter set. If we are interested in behavior under several variants of a game, then we have a tradeoff between statistical power and the number of variants that we can test.

Choosing a design means weighing concerns over obtaining potentially spurious effects against using less powerful tests. Opinions on this issue vary across the experimental community. We ourselves tend to prefer between designs whenever these are practical, as we believe these represent more conservative tests and we would rather err on the side of caution. Nevertheless, one must consider the context when making this design choice.

A large field within experimental economics deals with the evaluation of utility theories. These theories are formulated to describe individual responses to different choices. Given that, we might be quick to decide that a between-designed experiment that evaluates a theory about utility is unnatural. If individual A is risk averse over gains, while individual B is risk seeking over losses, could we really conclude that individuals in general have mirrored preferences?

Kahneman and Tversky (1979) demonstrate multiple failures of expected utility theory from questionnaire data. Using a between design, they expose some to gambles over gains and other to gambles over losses. They observe that risk-averse preferences on positive prospects are mirrored by risk-seeking preferences on negative prospects. They call this the reflection effect.

Soon after this, Hershey and Schoemaker (1980) criticized the between results of Kahneman and Tversky (1979) on the basis that a between analysis does not accurately represent a test of expected utility theory, because no individual preference reversals are occurring. They use the results from a within-designed experiment to claim that the preferences demonstrated are not consistent with reflectivity and thus prospect theory. However, they make multiple serious mistakes with the design. In two out of three treatments, all of the loss questions were presented before all of the gain questions, while in the third treatment they were presented side-by-side "to emphasize the experiment's focus on reflectivity." Furthermore, the order

of their questions within each section was never varied, nor was the order of the options always presented. Clearly, there is ample room for the biases discussed earlier to influence results here.

Budescu and Weiss (1987) use a within analysis of the same issue, but take into account all of the factors ignored by Hershey and Schoemaker (1980). They randomize the order and presentation of their gambles. They also use irrelevant gambles within lab sessions to try and minimize salience of earlier choices. These irrelevant gambles were also varied across treatments to ensure that they were not an additional source of bias. Their results support the original Kahneman and Tversky (1979) finding.

The main lesson here is that achieving proper identification can often be more important than providing an exact test of theory. While between analyses can be theoretically less palatable, we should remember that random assignment is a powerful tool that we may need to trust to produce useable results. Furthermore, this issue again demonstrates the importance of addressing potential sources of bias introduced by a within design. That said, Budescu and Weiss (1987) affirm that, with careful and clever design, one can access their statistical and theoretical advantages.

3. Simple examples: WTP elicitations

In terms of examples, one basic experimental setting in which the issue of choosing a within or a between design arises is a willingness to pay (WTP) elicitation. A researcher may ask her participants how much they would be willing to pay for a sandwich in their neighborhood bakery, and then how much they would be willing to pay for the same sandwich in the airport. Instead, the researcher could ask half of the participants how much they would pay at the bakery and the other half how much they would pay at the airport. Just laying out this simple experiment makes clear an immediate attraction of a within-subject design: here the experimenter gets twice as much data with the same number of individuals. Also immediately apparent is the fact that the experimental environments of the two methods are fundamentally different, because regardless of the order that the questions are asked in the within analysis, subjects have a reference or comparison point when responding to the second question. Since an experimenter cannot un-ask it in order to reset the individual to a resting state, unwanted psychological sources of variation are introduced once any question is asked.

An early argument in this spirit is made in Grice (1966), who criticizes the common use of surveys and within experiments in psychological studies for non-independence of questions and tasks. Poulton (1973) specifically criticizes within studies for ignoring what he calls range effects. This refers to the fact that exposure to a range of values in the lab affects subjects by lending contextual comparison to all scenarios other than the first. In a methodological paper, Greenwald (1976) criticizes within designs based on the effects of practice, sensitization and carry-over that confound causality. He outlines when within designs are problematic, mainly as a function of the type of question being asked by the researcher. All these papers argue that one should avoid these designs when the experimenter is interested in behavior in the absence of practice, when exposure to multiple treatments makes the individual overly sensitive to variations between the treatments, and when treatments have persistent effects.

If we wish to use a within design, we need to understand that exposure to multiple scenarios has psychological consequences. However, the fixes may not always be obvious. Sticking with the example of WTP elicitation, imagine that in the bakery/airport experiment described above, we vary the order of the scenarios presented to each individual in the within study, but their elicited value under the second scenario is always biased by their exposure to the first. Pooling across individuals exposed to the bakery first, we have a good measure of the bakery value and a bad measure of the airport value. The opposite is true for individuals exposed to the airport first. The natural thing to do would be to throw away the "dirty" measures and just compare the clean ones, but now we are back to square one with a between analysis. If we average the dirty and clean measures and looked at the difference between the two elicited values, we would need to maintain the assumption that the biases are of the same size and direction (i.e. the bias is independent of the scenario).

Kahneman and Ritov (1994) goal was establishing WTP (among other things) for an assortment of public goods. They use a survey that presents individuals with a headline and then asks for a response. Individuals repeat this process for a number of headlines. Worried that carry-over and range effects confound their within analysis, they analyze the correlation between question number and response within each individual. They find no substantial correlation and thus conclude independence of responses. This is a good simple example of problem recognition and response. But this is not a perfect fix to the problem. For example, Frederick and Fischhoff (1998) measured WTP for a variety of goods with within and between designs. The quantities of each good they were rating were varied. In both designs, individuals got the direction of the difference in WTP correct between the low and high quantities, but it was much bigger (by a factor of 2.5) in the within design. The authors suggest that in the within design, subjects feel more compelled to differentiate their answers by observing both scenarios at once and having to contrast them.

A good example of why these biases exist comes from the literature on evaluability. When we consider how much we value one product in isolation, we think about the amount of enjoyment we will receive when we consume it. When we consider whether or not we value one product more or less than another, we need to compare the products directly. The literature on evaluability focuses on the fact that when we are forced to make direct comparisons between products, there may be features of the products that are very easy to compare (think speaker wattage or thread count of sheets), and features that are not (think aesthetic design of speakers or color matching with sheets). If we overweight easily evaluable characteristics when we have to compare two products, decision under joint evaluation of products could diverge from decisions under separate evaluation. The application of this idea to the within versus between paradigm is straightforward; because within

designs necessitate exposure to more than one product (in the case of WTP elicitation), individuals could be using different criteria to supply their WTPs in within designs than in between designs, where they evaluate only one product in isolation. This line of reasoning can be extended to countless scenarios other than the WTP that we deal with in the laboratory. These points are made convincingly in Hsee (1996) who lays out the differences between joint and separate evaluation modes for consumption.

Further support for this is presented in Hsee and Leclerc (1998), who use a between designed experiment, in which subjects are exposed to one product, another product or both, they demonstrate that joint evaluation of products leads to different valuations than separate evaluations. This is a clear demonstration of one of the major concerns about within analyses (see also Hsee and Zhang, 2004).

The main lesson from this set of studies is that failing to take into account the complexities of a within analysis can make or break the validity of a result. If the order of asking a question matters significantly, then something other than experimental environment is contributing to the variation across groups. This variation could be from contextual referencing, learning, sensitization to changes, carry-over effects, or other psychological factors. If this bias is specific to one order but not the other, or specific to the orders in different way, then just varying the order of the questions might not be enough to remedy the problem. Considering that one of the three main advantages behind a within design is economizing on subjects, this is worth bearing in mind at the planning stage. Other tasks could be used to help "reset" an individual, time could elapse between elicitations or the experiment could be conducted in a segmented way. The goal should be to achieve an independent evaluation of each scenario by participants, and if a strong presence of the biases discussed is likely, a between approach may be preferable.

Another potential source of difficulty in WTP elicitation is what environmental economists have termed scope insensitivity. Kahneman and Knetsch (1992) and Desvousges et al. (1993) popularized the notion that when the contingent valuations of public goods are obtained between individuals, altering the quantity provided has little effect on WTP, directly contradicting basic economic theory. Within elicitations have not found this result (e.g., Brookshire et al., 1976). This would seem to indicate that we have to weigh the potential difficulty that individuals have in accurately perceiving the scope of goods in a between design, and forcing individuals to perceive scope differences in a within design. For farther discussion, and some limits of scope neglect, see Carson et al. (2001).

We believe that the important lesson to be learned from the environmental literature is that while it is easy for a researcher to look at two versions of a survey and see how they differ, respondents in a between framework see their information in a vacuum. Especially considering that the effort that individuals exert in responding can be difficult and costly to control, detecting responses to relatively minor changes can be difficult-to-impossible in a between design. In these circumstances, we would expect different results from the different design types, and which result is more meaningful likely depends on the context.

4. Different methods, different results

Experimentalists have recognized for a while that the framing of decisions can influence choices (Tversky and Kahneman, 1986; Andreoni, 1995; Epley and Gneezy, 2007, etc.). Framing decisions in the lab as contextual comparisons (as in within designs), or judgments made in isolation (as in between designs) can produce different results. Carry-over between scenarios can create patterns that would not exist in an isolated situation, or over-sensitivity to changes in parameters can develop that leads to observed differences where they would not otherwise exist.

The carry-over, context and sensitization effects mentioned earlier do not have natural tendencies to produce specific behavioral responses; their effects are functions of the circumstances. Experimenter demand effects however, may well have the tendency to magnify differences between evaluations. The act of moving a participant from scenario A–B makes them explicitly aware of the change to their environment. Often in modern experiments, the context in scenarios A and B are identical except for one parameter, to which the participant naturally pays attention. In an otherwise sterile and unchanging laboratory environment, participants might ask themselves how they should change their behavior in response, rather than first asking should they change their behavior. To the participant, it may seem as if the experimental variation is prompting a behavioral change, hence the label: experimenter demand. As discussed above, this concern was raised in the psychological literature over thirty years ago, in relation to the problems associated with individuals hypothesizing about experimenter intentions. This argument states that the kinds of variation we need to perform in experiments may result in decisions that do not represent natural preferences because the manipulation itself is unnatural.

With the goal of studying the power of the availability heuristic in determining probabilistic judgments, Milburn (1978) had participants fill out surveys estimating the likelihood of future events. Two of three groups are asked to estimate the probability of a series of future events, while each member of the third group estimates the probability of a single event. In surveys that elicit multiple probabilities, clear order-response patterns emerge. The probability of positive events occurring steadily increased over the time horizon while the probability of negative events steadily decreased. The between results were different however; the probability of positive events occurring decreased with time, in concurrence with the availability heuristic hypothesis. It seems likely in this case that contextual comparisons (carry-over and range effects) in the within surveys were influencing the results.

Fox and Tverksy (1995) studied ambiguity aversion. One group of gambles was clear with respect to the odds of winning, while the other was vague. Mindful of within-between differences, they gave some individuals the clear gamble, some the

vague gamble and some got both. The within analysis indicated ambiguity averse behavior while the between did not. The authors suggest that the comparative context in a within design is likely the cause for this result, arguing that individual analyses of gambles could be different when participants are considering two and asked to choose one than just being asked whether or not to take a gamble. When the gambles are presented together, one easy comparison that the individual can make between the two is that they know the odds for one but not the other.

In Gneezy (2005) participants were asked to evaluate the behavior of a car salesman who lies about the condition of the vehicle. The cost of his lie (repair costs he does not tell the buyer about) is either low or high. Some individuals are exposed to both conditions while other only see one of them. The within and between results both indicate that individuals consider the behavior less fair when the cost of the lie is increased, but individuals who were exposed to both scenarios changed their opinions drastically. In the between design, 36 percent of subjects called the behavior very unfair in the low cost scenario, while 62 percent of subjects called it very unfair in the high cost scenario. In the within design, 18 percent of subjects called it very unfair under low cost, but 68 percent called it very unfair under high cost. The percentage-point difference in rates is almost double with the within design. Gneezy (2005) postulates that the comparative context of the within analysis (and possibly also an experimenter demand effect) induces this difference.

The main lesson from this is that in a within analysis with a series of questions, we can analyze order-response correlations to get an idea of whether questions were answered independently.[1] This can be done if the order of the questions is varied (or if the order and questions are designed very carefully to avoid response trends). Otherwise, we cannot distinguish between carry-over bias and changing preferences. Perhaps a more serious problem with a within design is experimenter demand. It seems advisable to strive to change the scenario in a way that does not trigger change for change's sake. This could be expressed as an independence condition between the set of possible behaviors (in each scenario) and the set of scenarios to which an individual is exposed.

5. Different methods, similar results

An example of the within and between methods agreeing with one another is the experimental evaluation of eyewitness accuracy and confidence. The examples here represent the simplest of within and between designs, but this lends itself to an intuitive understanding of the procedures. All subjects participate in many rounds of judgments, but by limiting analysis to one question at a time, experimenters can still look at between differences despite the structure. There are potential order issues with such a design. For an optimal between analysis of specific questions, individuals should be presented with the questions in the same order. If two individuals see the same question at a different point in their sessions, directly comparing them is problematic unless we have a strong belief in independence. However, for an optimal within analysis, the question order needs to be varied to help minimize the effect of carry-over and sensitization biases.

In the context of skill-based, rather than preference based experiments, these concerns may not present such substantial obstacles. Subjects are aware of the fact that experimenters are interested in their ability, but this should not change the fact that subjects should be motivated to perform as well as possible. A systematic bias in these cases seems unlikely.

A paper demonstrating this is Deffenbacher et al. (1981). Participants were tested on their memory for faces and the confidence with which they recognized them. All subjects were shown a series of 50 faces and then asked to recall them a week later. To compute the correlation between accurate recognition and eyewitness confidence, the authors took two approaches. They first calculated the correlation coefficients for each question separately (between individuals) before averaging across questions, and second, for each individual separately (between questions) before averaging across individuals. The between method results in a coefficient of 0.48, while the within method gives 0.31. Smith et al. (1989) applied similar methods, and their results yield a between confidence-accuracy correlation of 0.14 and a within confidence-accuracy correlation of 0.17.

These results demonstrate that it is possible to design an experiment with both within and between tests, but once we switch to a multi-period format, there is a tradeoff we need to think about regarding question order. The particular experiments mentioned above appear to be good examples of lab tasks that are naturally closer to independent across periods than some of the examples used earlier. There are certainly stories that could be told about why this is not the case, but we see very similar results from both methods, especially in the second paper. The skill-based nature of the tasks would seem to make potential experimenter demand effects less likely.

6. Econometric considerations

When considering a design for experimental research, it seems prudent to consider econometric concerns when making choices. In this section, we discuss some general issues in this regard, as well as some specific examples. As an example, consider for a moment an experiment with multiple periods, where individuals can answer a series of questions, repeat tasks or play games with one or more other participants. The within-between distinction becomes murkier when we talk about longer horizons. The tasks can change over time or remain static, individuals can play with anonymous or known opponents, and opponents can be selected randomly or deterministically.

[1] We wish to emphasize that this is more of a check on the results rather than a fix of the problem.

More concretely, suppose that an experimenter is interested in performance with respect to a task. Individuals perform the task in either low- or high-stakes conditions. The nature of the research question requires multiple periods (e.g., because of learning). One option would be to split the sample into a high treatment and a low treatment and run both for a number of periods. Comparing the two treatments gives a between estimate of the behavioral difference. An alternative approach would be to have two treatments play for both high and low stakes, but to reverse the order of exposure across treatments. Treatment A might play low stakes for five periods and then high stakes for five, while treatment B would do the opposite. By analyzing how individuals change their behavior in response to a change in stakes, we can obtain a within estimate of the treatment effect (see Isaac and Walker, 1988a, 1988b).[2] The accuracy of this approach depends on whether any order biases cancel one another out across the two orders, as mentioned in the WTP section.

We have discussed the issues surrounding these approaches already. The face-recognition experiments from the psychology literature are examples of identical individual tasks with repetitions. However, simple correlation tests are inconsistent in the within-design case because of the standard omitted individual-heterogeneity issue when we talk about panel data. In fact, whenever we switch to a within design in a multi-period experiment, there are more serious econometric concerns that necessitate attention. The resulting panel data is not simply separable along treatment lines. By using a random-effects framework, we can (at least in principle) achieve more efficient results. By using a fixed-effects framework, we can obtain potentially efficient and consistent within estimates. In cases where the right-side variable is not randomly assigned across individuals (confidence or accuracy in the face recognition case), fixed effects must be used to achieve consistency.[3] It is important to remember though, that for ease and simplicity there should always be a way to design an experiment to obtain a consistent between estimate.

There are issues surrounding the move to a full-fledged effects model when analyzing experimental data. Critics argue that the correlation structure can never be fully untangled in cases of multi-player, multi-period games.[4] However, we argue that while not ideal, these approaches may be the best option in certain scenarios. Consider the case where the parameter set to be tested is too large for a between design to be feasible. We must then use a within design in which there is individual specific variation in the parameter value. Akin to varying the order of questions or treatments, one can parameterize each game using a random draw from the parameter set. Randomizing here alleviates concerns with order effects from a more structured design. Opponents can be randomly and anonymously re-matched after each round. We can consistently estimate the effect of the parameter on game behavior in this case with a random-effects model, subject to a couple of caveats.

The re-matching itself should not present a substantial issue unless the group size is very small. Playing the same opponent in adjacent rounds is unlikely (and is often explicitly prohibited), and combined with anonymity, it should reduce an individual's consideration of the possibility to a minor probabilistic contingency.

A problem with this approach is that when an individual is exposed to the play of others, she learns about the laboratory population. If this does not occur in a standard way across individuals, a time-period dummy in the effects model will not fully account for these effects. Similarly, if exposure to certain types of play creates emotional responses in certain individuals that can carry over into future periods, even including individual, partner and time effects would not account for these individual cross time effects.

An example of a paper that demonstrates this is Fehr and Gächter (2000), who experimentally test punishment and cooperation in public-goods games. Individuals played in groups of four for 24 rounds, either with or without punishment. Given the random and anonymous re-matching (in the "stranger" condition), the authors calculate their probability of having an identical group twice in a row as slightly less than five percent, and run a comparison stranger treatment where this probability was forced to be zero. Their results did not differ significantly, lending support to the random procedure. Fehr and Gächter (2000) perform their analysis in a number of ways. First, they take a between approach. They find contribution is much higher in the treatment with punishment, and a downward trend of contribution exists without punishment, but does not exist with punishment. They find that free-riding emerges as a focal point without punishment, but that it does not with punishment. Notice that they use a between approach to observe time-specific trends as well. To figure out why punishment increases contributions, they take an effects-model approach. They model the data as a panel, including time-period effects and group-effects in the stranger treatment. Their dependent variable is whether or not an individual was punished, and they include other's average contribution, own negative deviation and own positive deviation on the right hand side. Since the variables are definitely correlated with group effects, they take a fixed-effects approach. They find a substantial correlation between own negative deviation and a punishment outcome.

[2] Note there are two within options in this case. One could use individual differences between the two phases of the game, or average across individuals and use the overall group difference between the two phases. Both are within estimates; one is at the individual level and the other at the group level. With respect to individuals, the within-group approach combines within and between variation.

[3] The different types of effects models have intuitive interpretations that run parallel to the study of within versus between experiment designs. In a panel model, the between estimate uses data with individual-specific variation averaged out. Within estimates come from data with individual fixed characteristics differenced out, with any between-individual variation removed. The estimators are in many ways analogous to enforcing a between or within design ex-post on the data. A random, effects estimate is a linear combination of both the between and within estimate. A fixed, effects estimate is just the within estimate.

[4] In fact, the tradition among experimental economists has long favored non-parametric statistical analysis, with the most conservative wing considering that each session in which there is interaction among the participants constitutes only one independent observation.

Andreoni and Samuelson (2006) study the effect of a parameter in determining behavior in a twice-repeated prisoners' dilemma. They have 11 values to test, so a between approach is out of the question. They have individuals play for 20 rounds, randomly and anonymously matched with one other person each round. Before each game, one of the eleven parameter values is randomly chosen by the computer. The authors take a panel approach, using time and individual fixed effects.[5] In this case, the panel approach gave them the statistical power to test a hypothesis that would have been essentially impossible otherwise.

7. Conclusion

The methodological issue of within versus between designs is ubiquitous in experimental work. Between designs are more conservative in nature, but have limitations in some cases, while within designs have more power but potentially suffer from confounds. It is important to point out that researchers can combine the two designs in simple ways to access the advantages of both methods. A population of 400 individuals can be split into two treatments, A and B for 200 observations in each cell of a between-subjects design. However, if group A is asked question A and then question B and vice versa for group B, Now, the researcher has 400 observations in a within design with order effects controlled for and two between comparisons with 200 observation in each cell.[6] Whether or not all of the data can be used depends on the results, but this design provides double the amount of information given the sample size, hedged against a guarantee that at least one key comparison will be valid.

We have attempted to clearly delineate the issues surrounding this choice of methods. We provide example of experiments that use within, between or both designs; sometimes these approaches yield the same results and sometimes they do not. The issue is not a simple one and the choice depends on a number of factors. We hope to further the discussion and the development of a framework informing experimental researchers about the benefits and drawbacks of each approach, and have also discussed how one might minimize concerns of confounds (such as providing spatial or temporal distance between or among within decisions). In addition, we have discussed which scenarios are particularly susceptible to different behavioral patterns across the two approaches.

In general, we prefer between designs, but recognize the limitations involved. There may be cases where a within design is the only practical way to go. However, we believe that, although within designs look attractive, the researchers need to make the case that the confounds discussed above do not pose a challenge for the results. Of course, this article is not a final word on this topic. But there has been little explicit discussion of these issues within the community of experimental economics, and it seems time to begin to remedy this omission.

References

Andreoni, J., 1995. Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. The Quarterly Journal of Economics 110, 1–21.
Andreoni, J., Samuelson, L., 2006. Building rational cooperation. Journal of Economic Theory 127, 117–154.
Brookshire, D.B., Ives, B.C., Schulze, W.D., 1976. The valuation of aesthetic preferences. Journal of Environmental Economics and Management 3, 325–346.
Budescu, D., Weiss, W., 1987. Reflection of transitive and intransitive preferences: a test of prospect theory. Organizational Behavior and Human Decision Processes 39, 184–202.
Carson, R.T., Flores, N.E., Meade, N.F., 2001. Contingent valuation: controversies and evidence. Environmental and Resource Economics 19, 173–210.
Deffenbacher, K., Leu, J., Brown, E., 1981. Memory for faces: testing method, encoding strategy, and confidence. The American Journal of Psychology 94, 13–26.
Desvousges, W.H., Johnson, F.R., Dunford, R.W., Boyle, K.J., Hudson, S.P., Wilson, K.N., 1993. Measuring natural resource damages with contingent valuation: tests of validity and reliability. In: Hausman, J. (Ed.), Contingent Valuation: A Critical Assessment. North Holland, Amsterdam, pp. 91–164.
Epley, N., Gneezy, A., 2007. The framing of financial windfalls and implications for public policy. Journal of Socio-Economics 36, 36–47.
Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. The American Economic Review 90, 980–994.
Fox, C., Tverksy, A., 1995. Ambiguity aversion and comparative ignorance. The Quarterly Journal of Economics 110, 585–603.
Frederick, S., Fischhoff, B., 1998. Scope (in)sensitivity in elicited valuations. Risk Decision and Policy 3, 109–123.
Gneezy, U., 2005. The role of consequences. The American Economic Review 95, 384–394.
Greenwald, A., 1976. Within-subjects designs: to use or not to use. Psychological Bulletin 83, 314–320.
Grice, G., 1966. Dependence of empirical laws upon the source of experimental variation. Psychological Bulletin 66, 488–498.
Hershey, J., Schoemaker, P., 1980. Prospect theory's reflection hypothesis: a critical examination. Organizational Behavior and Human Performance 25, 395–418.
Hsee, C., 1996. The evaluability hypothesis: and explanation for preference reversals between joint and separate evaluations of alternatives. Organizational Behavior and Human Decision Processes 67, 247–257.
Hsee, C., Leclerc, F., 1998. Will products look more attractive when presented separately or together? The Journal of Consumer Research 25, 175–186.
Hsee, C., Zhang, J., 2004. Distinction bias: misprediction and mischoice due to joint evaluation. Journal of Personality and Social Psychology 86, 680–695.
Isaac, R.M., Walker, J.M., 1988a. Group size effects in public goods provision: the voluntary contributions mechanism. Quarterly Journal of Economics 103, 179–199.
Isaac, R.M., Walker, J.M., 1988b. Communication and free-riding behavior: the voluntary contribution mechanism. Economic Inquiry 26, 585–608.
Kahneman, D., Knetsch, J.L., 1992. Valuing public goods: the purchase of moral satisfaction. Journal of Environmental Economics and Management 22, 50–57.
Kahneman, D., Ritov, I., 1994. Determinants of stated willingness to pay for public goods: a study in the headline model. Journal of Risk and Uncertainty 9, 5–38.

[5] They also could have taken a random-effects approach, since the parameter at issue is randomly determined.
[6] Of course, the trade-offs might also depend on the access to a sufficient number of subjects.

Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47, 263–292.
Milburn, M., 1978. Sources of bias in the prediction of future events. Organizational Behavior and Human Performance 21, 17–26.
Poulton, E., 1973. Unwanted range effects from using within-subjects experimental designs. Psychological Bulletin 81, 201–203.
Rosenthal, R., 1976. Experimenter Effects in Behavioral Research, 2nd ed. Wiley, New York.
Smith, V., Kassin, S., Ellsworth, P., 1989. Eyewitness accuracy and confidence: within-versus between-subjects correlations. Journal of Applied Psychology 74, 356–359.
Tversky, A., Kahneman, D., 1986. Rational choice and the framing of decisions. The Journal of Business 59, S251–S278.
White, R.A., 1977. The influence of the experimenter motivation, attitudes and methods of handling subjects in psi test results. In: Wolman, B.B. (Ed.), Handbook of Parapsychology. Van Nostrand Reinhold, New York, NY, pp. 273–304.