Exp Econ (2012) 15:485–498
DOI 10.1007/s10683-011-9309-1

Session-effects in the laboratory

Guillaume R. Fréchette

Received: 20 April 2011 / Accepted: 25 October 2011 / Published online: 11 November 2011

Abstract In experimental economics, where subjects participate in different sessions, observations across subjects of a given session might exhibit more correlation than observations across subjects in different sessions. The main goal of this paper is to clarify what are session-effects: what can cause them, what forms they can take, and what are the potential problems. It will be shown that standard solutions are at times inadequate, and that their properties are sometimes misunderstood.

Keywords Session-effects · Data analysis · Hypothesis tests · Experimental design · Unit of observation · Matching procedure

1 Introduction

In experiments, where subjects in the same treatment are usually separated in a number of individual sessions, there may be correlations between observations of subjects who participated in the same session. Similarly, when analyzing data from multiple members of a family, observations from siblings might exhibit more correlation than those from individuals across households. In experimental economics this is sometimes referred to as the session-effects problem. To date, there is no explicit articulation of this problem, thus making it difficult to know if a given solution is appropriate. Nonetheless, the session-effects problem has become important in experimental

I wish to thank the editor and two anonymous referees, Marina Agranov, Jeffrey Carpenter, Gary Charness, Yan Chen, Vincent Crawford, Pedro Dal Bó, Mark Dean, John Duffy, Martin Dufwenberg, Drew Fudenberg, John Ham, Burkhard Schipper, Andrew Schotter, Chloé Tergiman, Kevin Thom, Isabel Treviño, and Emanuel Vespa for helpful comments and the NSF via grants SES-0519045, SES-0721111, and SES-0924780, and the Center for Experimental Social Science (CESS) for their financial support.

G.R. Fréchette (✉)
New York University, New York, USA
e-mail: frechette@nyu.edu

economics in the sense that it influences how data analysis is performed, how experiments are designed, and what questions can be asked. Given the increasingly widespread use of experimental techniques and the fact that many non-experimenters now rely on experimental results, issues central to experimental methods and the way data analysis is performed are of more general interest.

Unfortunately, since there is no clear articulation of the problem, it is difficult to formulate an appropriate response to it. The present paper has two aims. First, to formulate more clearly what session-effects are and how they could arise. In particular two different types of session-effects will be identified: static and dynamic session-effects. Second, to explore the implications for experimental design and data analysis. Note that the focus of this paper is not on how to deal with session-effects. Rather, the aim is to define more clearly the issue. This, in turn, will help to evaluate our current practices and to provide a basis for future work.

Without having defined session-effects precisely, it is nonetheless clear that correlation in subgroups of a given population are frequent in many other areas beside experimental economics. For instance analysis of survey data involves clustering issues (Sakata 2002). Experimental data in ecology (Warton and Hudson 2004), biology (Williams 2000), medicine (Altman and Bland 1997) and other areas have repeated measurements per treatment which induce similar correlations, and the analysis of genes and their comparison across subpopulations generates the same type of problems (Excoffier et al. 1992). Each of these fields has developed different, although often closely related, methods for dealing with these problems given the particular details of their applications.

The problem of session-effects is often addressed in one of the two following ways in experimental economics.[1] One solution is to use session averages of the variable of interest. The intuition for why this could eliminate the problem is fairly straightforward. Imagine a data set of observations from multiple subjects divided in sessions. Many statistical tests assume that the observations are independent, an assumption that is violated if there are session-effects. If the session averages are treated as the unit of observations, then the correlation that existed because of the session-effects is believed to be no longer present.

The other solution is not to replicate the task of interest in an experimental session, i.e. play the game or have subjects make the relevant decision only once. Why these two methods are thought to resolve the problem will be discussed in more detail below. It should be clear that these solutions are not without costs. Both reduce the number of observations available (for a given number of sessions) and thus increase the actual cost of running experiments. These solutions also reduce the power of the statistical tests that can be performed. To illustrate this point, consider an experiment with two treatments, four sessions per treatment, 16 subjects per session, and each subject plays 10 repetitions of the same game (that's a total of 1280 observations, but only 8 sessions). For simplicity, imagine that the pooled standard deviation of the variable of interest is the same if each data point is treated separately or if the data is first averaged by session. In that case, the smallest treatment effect that would lead to a rejection of the null hypothesis that there is no treatment effect needs to be

[1] See for instance Davis and Holt (1993, pp. 527–528) and Friedman and Sunder (1994, pp. 98–99).

more than 15 times greater using a t-test at the 5% level if the data is first averaged by session as opposed to if each observation is used separately. Clearly, this is an important loss in power.

Furthermore, if the researcher believes that the behavior of interest is the one which occurs after the subjects have understood the game and that this is only possible through practice, then the second solution (having subjects take a single decision) is not possible. Thus if the question of interest requires controlling for observables, this could make it extremely difficult to ask some types of questions since averaging by session does not allow to keep track of subject specific variables. A simple example of this is how can one properly estimate a bidding function in an auction experiment using session averages if the height of the subject is a key determinant.[2] Also, if one is interested in the interaction between a variable and the treatment, then one is forced to study it within the second setting (a one-period experiment) by introducing variation in the variable of interest within each session. This constrains the experimenter to using cross-sectional analysis and thus limits his ability to control for other factors which might be relevant. Moreover, it will be shown that both methods can create new problems and that depending on the source of the session-effects, they may not even address the problem they are meant to solve.

2 The problem defined

The session-effect problem is defined as a within session correlation in the variable of interest (or the residual) once the relevant factors are controlled for.[3] This could result either from some relevant factors being unobservable (for example a hormone that cannot be easily measured), or from the fact that the researcher is ignorant of some relevant factors which could be controlled for if their importance was known (for example the gender of the person conducting the experiment). Typically we think of situations where the problem is a positive within session correlation. Thus, the greater the session-effect problem, the lower the variance in the variable of interest within a session relative to the total variance.

Suppose the variable of interest is y, which is determined by X, ε, and potentially T; where X are observable factors known to determine y whereas ε are either not known to matter by the experimenter or not observable. T takes value one for the treated sessions and zero for the control, and the goal of the experiment is to determine if T has an impact on y, that is the question asked is if $E(y \mid T = 1) - E(y \mid T = 0) \equiv \theta \neq 0$.[4] However, the experimenter might also be interested in estimating the marginal effect of some observable $\frac{\partial y}{\partial x_k} \equiv \beta_k$ as in auction experiments where one is interested in estimating the bidding function. Absent

[2] If the variables that need to be controlled for only take a small set of values, the sample can simply be separated into subgroups, but when controlling for continuous variables this is not an option.

[3] Clearly "relevant" here excludes the source of the session-effects which although relevant, are either unobserved or unknown to matter to the experimenter.

[4] We assume that there is only one treatment for simplicity of exposition, but nothing in this analysis depends on this assumption.

session-effects, the typical assumption is that

$$Cov\left(y_{ips}y_{jqt} \mid X_{ips}, T_{ips}, X_{jqt}, T_{jqt}\right) = 0 \quad \forall p, q, s, t \text{ and for } i \neq j, \tag{1}$$

where the subscript indicates the subject, the period, and the session.[5] Session-effects are present if (1) above is true for $s \neq t$ but not for $s = t$, that is if

$$E\left(\varepsilon_{ips}\varepsilon_{jqs}\right) \neq 0. \tag{2}$$

In experiments, identification of θ (or of β) is typically done through randomization and control of observables using either measurement or design (within subject designs and dual market procedures). In order to understand the role of (1), it is easier to first simplify the environment further. Imagine a setting where there is only one observation per subject so that we do not need to keep track of the periods. Hence, we have a cross-sectional data set. Denote the covariance matrix of the error term by Ω (or $E(\varepsilon\varepsilon') = \Omega$). A standard assumption is that $\Omega = \sigma^2 I$, where I is the identity matrix. In other words, one assumes spherical errors, that is that the diagonal elements are the same (homoscedasticity) and that the off diagonal elements are 0 (noautocorrelation). This is the assumption under which the variance of an estimator is typically derived. For instance, for OLS, the Variance of the estimator (where $Z = [X, T]$) is given by $(Z'Z)^{-1}Z'\Omega Z(Z'Z)^{-1}$ and thus under the assumption of spherical errors, one obtains that $Var(u) = \sigma^2(Z'Z)^{-1}$. Equation (2) (session-effects) is a type of autocorrelation, in other words it implies that Ω is not diagonal. Putting aside session-effects for a moment, violations of (1) for our simplified case of one observation per subject would occur if there was a correlation (unaccounted for) between the observations of different subjects. This could happen, for instance, if proper experimental control is not used and some subjects are allowed to communicate and influence each other's decisions. This might be a greater concern for online experiments performed in a small group (e.g. in university dormitories). Coming back to the more general case where the experimenter records multiple observations per subject, autocorrelation becomes a clear possibility since in many cases it seems plausible that an individual's observations are correlated. That is why (1) is stated to allow for such correlations (see also footnote 5).

If they are ignored, session-effects produce two principal problems for data analysis. First, if the session-effects are such that $E([X, T]'\varepsilon) \neq 0$, in other words if the session-effects and the observables are correlated, then standard estimators (OLS for example) will be neither consistent nor unbiased. On the other hand, if session-effects are ignored but these are not correlated to the observables of interest, then although the appropriate estimators might be consistent and unbiased, the computed variance of the estimator will be incorrect. This, clearly, will affect hypothesis testing, and

[5] This specification is more restrictive than required, but this has no bearing on what follows. For instance one can easily accommodate sequential games where an agent's decisions depend on prior decisions.

Note also that (1) allows for correlation of the observations across the decisions of a subject. Although this is not addressed in the current paper, if this is relevant for a given application, it would need to be taken into account in the estimation. An example where this is discussed explicitly in an experimental context is Fréchette (2009).

specifically tests to determine if there is a treatment effect or not. In particular, given that most session-effects are likely to lead to within session positive correlation, this will lead to the computed variance to be lower than the true variance of the estimator, and thus to rejecting the null hypothesis too frequently. In such cases, there are two levels of concern. One is to use the correct variance for the estimator at hand, the other (less critical) concern would be to use estimators that are more efficient.

3 Potential sources

What could cause session-effects as in (2) above? One example sometimes mentioned is experimenter effects. Experimenter effects happen if y is affected by the person who conducts the experiment. This was an important consideration for instance in the design of Roth et al. (1991). In that experiment, behavior in two games is compared across four countries: the United States, Yugoslavia, Japan, and Israel. The experimenters in Yugoslavia, Japan, and Israel are all different; and thus any observed difference could be the result of the experimenter, as opposed to the population of the subjects. To address this, they had each experimenter conduct sessions in the US and used the US data to test for experimenter effects, which they did not find. There is evidence of experimenter effects in other fields however (see for instance Lewejohann et al. 2006). Hence, if an experiment is conducted by multiple experimenters, and there are experimenter effects which the experimenters are unaware off (and thus do not control for), this would lead to session-effects.

A very different example of session-effects is what happens in Median game experiments. Median games were studied by Van Huyck et al. (1991). In those games subjects choose any integers from 1 to 7 and the payoffs of each player depend on their own choice and the median of all choices, such that all subjects choosing the same number constitute a NE, and that is true for every number. The payoff matrix remains constant for the first 10 periods. In all 12 sessions: (1) it is never the case that all subjects make the same choice in period 1, (2) what is the median in period 1 varies across sessions, and (3) whatever is the median in period 1, it is the median in every period of that session. Furthermore, in 11 of the 12 sessions, every player makes the median choice in period 10. This clearly corresponds to the definition of session-effects since the behavior of subjects by the end is determined by the specifics of what happens in period 1 of their session.

Clearly these two examples are very different. In the first one, the session-effect is created by something that is constant throughout the session and is not a function of the behavior during the session. In the second example, the end result is a function of what happened during the session. It will be helpful to differentiate between these two types of session-effects, which I will call static session-effects and dynamic session-effects. Thus, static session-effects exhibit a within session correlation in the variable of interest (or the residual) once the relevant factors are controlled for, and the source of the correlation is constant within a session. On the other hand, dynamic session-effects also result in a within session correlation in the variable of interest, but in this

case it is caused by the realizations of a variable that varies over the course of an experiment.[6]

4 Implications

This section will present five claims often made about data analysis (either explicitly or implicitly) in experimental papers. The conditions under which they are correct or not will be established.

Myth 1 *Using session averages as the unit of observation eliminates the session-effects problem.*

Averaging the variable of interest by session ($\bar{y}_s$) and performing all statistical analysis using session averages as the unit of observation is a common occurrence in experimental papers and this is done to eliminate the problems caused by session-effects (see for instance Blume, Duffy, and Franco's paper in the 2009 *AER* (p. 1185) or Vanberg's *Econometrica* paper in 2008 (p. 1473)—Note that these and other references to methods of data analysis in experimental papers are simply provided to show that such approaches are used in recent years in leading journals. Every procedure highlighted in this paper has been used by myself in at least one paper.). There are two distinct ways in which this may not address the problem: if (1) the session-effects are correlated with the treatment or the variable of interest, or (2) if the session-effects are correlated to an unobservable that is either unknown or not observed and this unobservable has an impact across sessions.

Here is an example in an auction setting.[7] In this case the experimenter is trying to estimate the bidding function in a sealed bid first price auction (the impact of value on bid). This example is based on the observation by Chen et al. (2005) that women's bidding behavior is affected by where they are in their menstrual cycle, and since subjects tend to be undergraduate students who often live in the same dormitories, their cycles may exhibit positive correlation.[8] Thus, it could be that since different sessions are conducted on different days within the month, some sessions display more shaving of bids, all else being equal, than others conducted on different days (in a standard bidding function relating value to bid, this would imply variations in the coefficient that relates value to bid). Simply put, there is a correlation between when the sessions are conducted and how subjects bid because of where the women are in their menstrual cycle.[9] If the researcher is not aware of this relation between

[6] This could include the decisions of others in the session, the realization of a random variable, the past outcome of some decision, etc.

[7] Note that to use session averages as the unit of observation with auction data, one would typically have to transform the variable of interest. For instance, one could use the bid minus the value or the bid divided by the value (depending on the bid function).

[8] Menstrual Synchrony has been observed in many species, including humans, for females living together (see Stern and McClintock 1998 and the references therein).

[9] Examples of this sort can be constructed with any hormone which varies over a period of time in a similar way for all (or a subgroup of) subjects. For instance, there is evidence that testosterone increases

Table 1 Hypothetical data for Myths 1, 3, and 5

Treatment	0	0	0	0	1	1	1	1
Sessions	1	2	3	4	5	6	7	8
Tall RA	0	1	0	1	0	1	0	1
DGP	$y_{ips} = x_{ips} + 1\{\text{Tall RA}_s\} + \varepsilon_{ips}$							
$\overline{(y_{ips} - x_{ips})}_t =$	$\overline{\varepsilon}_1$	$\overline{\varepsilon}_2 + 1$	$\overline{\varepsilon}_3$	$\overline{\varepsilon}_4 + 1$	$\overline{\varepsilon}_5$	$\overline{\varepsilon}_6 + 1$	$\overline{\varepsilon}_7$	$\overline{\varepsilon}_8 + 1$

DGP: Data Generating Process

$\overline{(y_{ips} - x_{ips})}_t$ is the session average of $(y_{ips} - x_{ips})$

hormonal levels and bidding behavior then he or she cannot control for it. Simply taking session averages does not eliminate the problematic correlation.

To illustrate better the mechanics of the problem, I will consider a simpler example to which I will come back on a few occasions. Again, take the case of a second price auction, but this time the factor that affects bids is whether or not the research assistant conducting the experiment is tall or not. When he is tall, subjects overbid on average. The data generating process is given by $y_{ips} = x_{ips} + 1\{\text{Tall RA}_s\} + \varepsilon_{ips}$ where $1\{\text{Tall RA}_s\}$ takes value 1 if the RA is tall and 0 otherwise. Table 1 represents a hypothetical series of 8 sessions. The table indicates the per session average error term if the impact of the height of the RA is not known to the experimenter. It is obvious that the covariance of the per session average error terms for any two sessions that were conducted by the tall RA is not 0.[10]

To summarize, in both examples the error terms within a session and across sessions are not independent. In other words, the off-diagonal elements of the variance-covariance matrix of the error terms are not all zero, even after taking session-averages.

Myth 2 *Using (non-parametric tests on) session averages is a more conservative approach.*

It is often said, and written, that averaging observations at the session level and using non-parametric tests to identify treatment effects is conservative, and this is meant to imply that if the null hypothesis is not rejected using a parametric test on all the data, then it would not be rejected using non-parametric tests on averages

aggressiveness (see Beatty 1992 for evidence in rodents and Lehrer et al. 2004 for a brief summary of the evidence in humans) and it is known that testosterone decreases during the day (Dabbs 1990). Hence, any task where aggressiveness matters could result in sessions conducted at a similar time in the day being correlated. Similarly, Schipper (2011) finds that risk aversion correlates with testosterone and cortisol in men.

Note that I am not giving examples because I think they are likely, or important in magnitude, but rather they are provided as an illustration of something that is possible.

[10] In the particular example of Table 1 there is no correlation between the point estimate of the treatment effect. If instead the tall RA had conducted sessions 5 through 8 and the short RA sessions 1 through 4, then the session-effect would also be (perfectly) correlated to the treatment, which would lead to additional problems.

Table 2 Hypothetical data for Myth 2

Treatment	0		0		0		1		1		1	
Session	1		2		3		4		5		6	
y	1	7	2	8	3	9	4	10	5	11	6	12
Average	4		5		6		7		8		9	

of the data.[11] Note that even though using session averages does not imply using non-parametric tests, it usually does in practice because the sample size becomes so small that it seems unreasonable to rely on the central limit theorem. The underlying notion seems to be that since the variance of an estimator goes down as sample size grows, this approach must make it more difficult to reject the null. Here is a simple example where this is not true. Imagine that you have 6 sessions (3 per treatment) each with 2 data points and there are no session-effects and no treatment effect. The data is represented in Table 2. Using a ranksum test over session averages one would reject the null hypothesis that there is no treatment effect at the 10% level. On the other hand, using a t-test over the entire data the null hypothesis of no treatment effect would not be rejected. As this example makes clear, using a non-parametric test over session-averages is not always more conservative. Since the average reduces the sample in a very specific way, whether or not this is more conservative depends on the relation between the variance within the sessions versus across their means. The problem with using session averages is that our conclusions are not affected anymore by the variance within the session, which may, or may not, be more conservative depending on the specifics of the data.

Myth 3 *Experiments where the task is performed only once eliminate session-effects problems.*

The types of session-effects for which this will not provide a solution are static session-effects. Just as for Myth 1 above, if the session-effects are correlated to the variable of interest, then the estimator will not be consistent. Also, as for Myth 1 above, if the session-effects are correlated to a latent variable, then the variance of the estimator will not be estimated properly. Returning to the example presented in Table 1, imagine that each subject participated in only one auction, it is still the case that the covariance of the errors from subjects that participated in sessions conducted by the tall RA will display positive covariance. However, in this case (differently from the case of Myth 1), the problem will be relevant even if the latent variable is session specific but uncorrelated across sessions. Similarly, the first round or period of an experiment is not immune to static session-effects. Examples of papers where subjects play only once and where this is believed to eliminate session-effects abound. A typical example is given by Fischbacher et al. (2001) in which they write: "since all subjects played only once, all 44 decisions are independent observations." Charness et al. (2004) provide an example of incorrectly claiming that focussing on period one eliminates concerns for session-effects.

[11] For example both Seinen and Schram (2006) and Ivanova-Stenzel and Salmon (2008) refer to a pairwise Mann-Whitney test on session averages as conservative.

Myth 4 *A turnpike design eliminates session-effects problems.*

The turnpike design, also known as a zipper design, was introduced by Cooper et al. (1996) and it ensures that subjects cannot influence the behavior of future subjects they interact with.[12] To understand why a turnpike design may not eliminate session-effects, it will be helpful to understand why it could sometimes solve the problem. Take the case of a game played in pairs where subjects are randomly rematched across periods. A player, say player A, is imagining that he may change his behavior early to change the behavior of the subjects he will interact with in the future. This could happen if, for instance, he first interacts with player B. In period two players B and C play together, and in period three player A plays with C. Player A may decide to change his behavior to affect what B does with C next, in the hope that it will in turn change how C behaves once matched with A. If the experiment uses a turnpike design (and this is explained and understood by the subjects), then subject A would know this is not possible and thus would not change his behavior early to affect what happens latter.

As the example makes it clear, this does not provide a solution if the problem is static session-effects, and furthermore it does not necessarily resolve dynamic session-effects. Dynamic session-effects can have many causes. First, they could result from subjects updating their beliefs about some relevant population parameters or their beliefs about relevant parameter values in the subsample of the population with whom they are interacting. Second, it could result from subjects trying to influence what others do (an example of this would be strategic teaching as in Camerer et al. 2002). Third, it could arise because subjects (partly) imitate the behavior of others. For instance, some subjects may not be able to solve a game by themselves but nonetheless recognize the solution when they see someone else play it. Many other factors, such as reciprocity, can also generate dynamic session-effects. Clearly some of these sources of the problem, such as strategic teaching or reciprocity, can be eliminated by using a turnpike protocol. However, if dynamic session-effects arise because of imitation, then using a fixed pairing, random re-matching, a round robin procedure, or the turnpike protocol will not help.

Myth 5 *Separating sessions into smaller subgroups alleviates the session-effects problem by increasing the number of independent observations per session.*

This can be once again illustrated by the example of Table 1. Even if each session is divided into subgroups, the average $(y_{ips} - x_{ips})$ for each subgroup in sessions conducted by the tall RA will still be $\overline{\varepsilon} + 1$. Hence, the covariance of the subgroup averages of the errors will again be positively correlated for the subgroups in sessions

[12] In the turnpike protocol, subjects cannot influence the decisions of future subjects they will be paired with through the decisions they take in the current match. To understand this procedure, line up the subjects in two equal rows, each subject interacts with the person in front of them, then all the subjects in one of the row moves one seat, and interact with the new person in front of them (the person at the extremity moves back to the beginning of the line). This can be repeated until the players would end up in their original position, at which point it must stop. Note that this is different from what is know as the absolute stranger design which only requires that no subject is matched with someone else twice.

conducted by the tall RA. However, in this case, not only does it not solve problems caused by static session-effects of the types described as problematic for Myth 1, such an approach will lead to the standard variance computation for the estimator not to be correct for any types of static session-effects. Since static session-effects are shared in the entire session, dividing the session into subgroups necessarily results in correlation between the subgroups of a given session.

Moreover, if there are dynamic session-effects and those are due to subjects relying (partly) on imitation, or being influenced by feedback, then rematching in a small group only exacerbates such effects since the feedback always comes from this smaller group. This is illustrated with one last example. Imagine an experiment in which agents choose, in a two player game, between two options, A and B: when they believe that more than 60% of the other subjects will choose B, they prefer selecting B, otherwise they prefer choosing A. In the first period beliefs are distributed uniformly. In all other periods subjects look at the past decisions of the people they were matched with and average those decisions to form their beliefs (i.e. if they have played T periods so far and a of these choices were A's, they believe that A will be selected by someone with probability $\frac{a}{T}$). This means that a subject that was matched with someone who played B in period 1 will play B in period 2 (he now believes others play B with probability 1), and if the person he is matched with plays A in period 2, he will play A in period 3 (since he then assigns probability $1/2$ that others will play B).[13] Subjects are randomly rematched within their group every period and they play for 10 periods. The comparison involves either 10,000 randomly generated sessions with 16 subjects each versus the case where each session is divided into 4 subgroups. Both of these can be compared to what would happen in a large population. Clearly, in a large population such a process leads to all individuals selecting A since more subjects select A than B in period 1. This can be confirmed by performing a baseline simulation with a single session that includes 160,000 subjects (so that the sample size is the same as in the example). In that baseline simulation, the correlation between the choice in period 1 and period 10 is 0.00 and the frequency of B choices in period 10 is 0.00 (the frequency of B choices in all three simulations is 0.40 in period 1). In the case with sessions of 16 subjects all interacting with each other, the correlation between period 1 and period 10 choices is 0.07 and the frequency of B choices is 0.02 in period 10. On the other hand, if the sessions are divided in 4 subgroups, then the correlation between period 1 and period 10 choices increases to 0.33 and the rate of B choices in period 10 is 0.17. Two aspects of this example are informative. First, dividing the sessions into smaller subgroups leads to a higher correlation between the behavior at the end of the session and the behavior in period 1. Second, the decisions are further away from what happens in a large population. Loosely speaking, subdividing a session makes the process more sensitive to the specifics of the group. Note that the pattern highlighted in the example has nothing to do with the kind of learning assumed in the data generating process (the same could be obtained with other forms of learning, including Bayesian). It also does not require agents to

[13] This structure is in the spirit of fictitious play (see for instance Fudenberg and Levine 1999). However, for simplicity, the example, unlike in a typical model of fictitious play, assumes that all decisions after period 1 do not depend on the prior subjects hold in period 1.

engage in "learning" of that sort, imitation as well as many other processes whereby the actions of some subjects affect each other can be used to construct such examples.

Finally, if static session-effects are a problem, then when conducting sessions with a partner design (as opposed to stranger, turnpike or other random rematching scheme), averages of different pairs in a session are not independent. There are numerous examples of partner data being treated differently from stranger data (see for instance Croson 1996). There is also an increasing number of examples of random re-matching in fixed subgroups of a session such as in Charness et al. (2007) and Apesteguia et al. (2007).

5 Conclusion

Many of the methods typically used to address session-effects work only for certain types of session-effects. Because the process by which session-effects come about is typically not specified, this fact has not been noted before. In particular, if the session-effects are correlated with the variable of interest or if they are correlated to factors that are not controlled for, then neither using session averages nor performing the task only once eliminates the problem. Other approaches such as using a turnpike design work for certain dynamic session-effects but not if the dynamic session-effects are caused by belief updating about population parameters or if they are due to imitation. Furthermore, this does not work for static session-effects. Separating sessions into smaller groups, an increasingly common method, does not solve the issue if the source are static session-effects. In the case of dynamic session-effects it may actually amplify the size of the effect, although it would create independence between the subgroups.

From a statistical point of view, the optimal design is one which gives absolutely no feedback about the behavior of other subjects, has many periods, and many small sessions. This would eliminate the possibility of dynamic session-effects (by eliminating feedback) and allow the best chance to identify static session-effects (through the many repetitions and with the numerous sessions). Unfortunately, eliminating feedback could very well hinder learning (see for instance Armantier 2004). From an experimental point of view this type of design has many drawbacks.

One easily implemented solution in many situations is to use clustering at the session level in the computation of the variance-covariance matrix. One caveat is that the properties of such techniques in samples of the typical size in experiments are not well known.[14] Notwithstanding such concerns, the advantage of this method is to provide a robust approach to testing, however it is not efficient. One can gain more efficiency, at the expense of a less robust approach, by modelling the source of the session-effects more explicitly and estimating that model. If the problem is believed to be static session-effects, then estimators such as a fixed or random effects

[14] Cameron et al. (2008) show in specific environments, through the use of Monte Carlo simulations, that when there is a small number of clusters, other corrections (different from the one typically employed) have better performance. However, the results as to what corrections perform best may be sensitive to the specifics of the data generating process.

estimator will often be a sensible way to proceed. Note that even if one is dealing with static session-effects, using an unbiased and efficient estimator may sometimes be more complicated than simply using fixed or random effects. One such common occurrence are cases involving dynamic panel data sets (where the dependent variable depends on lagged values of itself). A useful estimator in some of these situations will be a correlated random effects estimator.[15] Other situations that go beyond fixed and random effects are cases where the session-effects are static but interact with other factors that determine the variable of interest. These could be modelled as variable (fixed or random) coefficients models.[16] When one is concerned with the presence of dynamic session-effects, then the appropriate estimator will necessarily depend on the specific source of the problem. As an example, if the issue is that subjects' behavior in a group is influenced by the feedback they receive about what others did in the previous period, then simply controlling for that could suffice. However, one may suspect the presence of dynamic session-effects without knowing the exact source or the specific form they take, in such cases simply clustering the standard errors may be a reasonable approach, it is not efficient, but more robust.

Recently there has been an increase in the number of laboratory experiments conducted in the field. Given that the conditions often involve less control and sometimes changes of location and environment, one could speculate that the potential for session-effects is greater. First, because the possibility for a group to interact outside of the game under study is greater (for instance it might be more difficult to refrain subjects from talking, but also they are more likely to interact together after the experiment). Second, since changes in location and environments could trigger reactions that the experimenter is unaware off. This highlights the importance of trying to minimize the variability in setting and protocol; and the importance of anonymity during the experiment (for instance paying subjects privately) in field settings. Another recent development is the increase in popularity of neuroeconomics experiments. In those, it is usually the case that only one subject is studied at a time. Thus, dynamic session-effects are a lesser concern. They could only occur if subjects react to outcomes in the experiment in a way that is unknown to the experimenter (as opposed to resulting from the interactions with other subjects). Static session-effects can still be problematic. In some cases however, those can be absorbed into a subject fixed (or random) effect.

In the process of finding examples for this paper, I became convinced that it is difficult to find credible (and potentially large) static session-effects except for cases that are so evident that a good experiment should have addressed them directly, such as in the Roth et al. (1991) paper. One exception to this statement might be gender. The gender composition of a group has been shown to affect behavior (Gneezy et al. 2003), but only in certain environments. Dynamic session-effects would seem like a potentially more common occurrence, but it is difficult to find many situations where they seem enormous, such as in median type games. This is not to say that session-effects should not be accounted for, however in many experiments it would seem

[15] See Chamberlain (1980) and Heckman (1981), or Wooldridge (2002) for a more recent discussion (under "initial conditions problem" and also "Chamberlain's random effects probit models").

[16] See for instance Hsiao (2003). See also Arellano (2003) for a reference on dynamic panel data analysis.

more important to account for the fact that one has repeated observations for each subject. The literature is replete with examples of heterogeneity in behavior across subjects: bargaining (Fréchette et al. 2005), auctions (Ham et al. 2005), and repeated prisoner's dilemma (Aoyagi and Fréchette 2009) just to name a few examples. Hence, it seems that we may want to allocate more attention to properly take those into account. It is true that clustering by session will allow for some forms of subject specific heterogeneity, but we might want to consider more direct and efficient ways to deal with those.

References

Altman, D. G., & Bland, J. M. (1997). Statistical notes: units of analysis. *British Medical Journal*, *314*, 1874.

Aoyagi, M., & Fréchette, G. R. (2009). Collusion as public monitoring becomes noisy: experimental evidence. *Journal of Economic Theory*, *144*(3), 1135–1165.

Apesteguia, J., Huck, S., & Oechssler, J. (2007). Imitation-theory and experimental evidence. *Journal of Economic Theory*, *136*, 217–235.

Arellano, M. (2003). *Panel data econometrics*, *OUP Catalogue*. London: Oxford University Press.

Armantier, O. (2004). Does observation influence learning? *Games and Economic Behavior*, *46*(2), 221–239.

Beatty, W. W. (1992). Gonadal hormones and sex differences in nonreproductive behaviors. In A. A. Gerall, H. Moltz, & I. L. Ward (Eds.), *Handbook of behavioral neurology: Vol. 2*. New York: Plenum Press.

Blume, A., Duffy, J., & Franco, A. M. (2009). Decentralized organizational learning: an experimental investigation. *American Economic Review*, *99*(4), 1178–1205.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2002). Sophisticated EWA learning and strategic teaching in repeated games. *Journal of Economic Theory*, *104*, 137–188.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, *90*, 414–427.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, *47*, 225–238.

Charness, G., Fréchette, G. R., & Kagel, J. H. (2004). How robust is laboratory gift exchange? *Experimental Economics*, *7*(2), 189–205.

Charness, G., Fréchette, G. R., & Qin, C.-Z. (2007). Endogenous transfers in the prisoner's dilemma game: an experimental test of cooperation and coordination. *Games and Economic Behavior*, *60*(2), 287–306.

Chen, Y., Katuscak, P., & Ozdenoren, E. (2005). *Why can't a woman bid more like a man*. Mimeo.

Cooper, R., DeJong, D. V., Forsythe, R., & Ross, T. W. (1996). Cooperation without reputation: experimental evidence from prisoner's dilemma games,. *Games and Economic Behavior*, *12*, 187–218.

Croson, R. (1996). Partners and strangers revisited. *Economics Letters*, *53*, 25–32.

Dabbs, James M. J. (1990). Salivary testosterone measurements: reliability across hours, days, and weeks. *Physiology & Behavior*, *48*(1), 83–86.

Davis, D. D., & Holt, C. A. (1993). *Experimental economics*. Princeton: Princeton University Press.

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, *131*, 479–491.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*, 397–404.

Fréchette, G. R. (2009). Learning in a multilateral bargaining experiments. *Journal of Econometrics*, *153*(2), 183–195.

Fréchette, G. R., Kagel, J. H., & Morelli, M. (2005). Behavioral identification in coalitional bargaining: an experimental analysis of demand bargaining and alternating offers. *Econometrica*, *73*(6), 1893–1938.

Friedman, D., & Sunder, S. (1994). *Experimental methods: a primer for economists*. Cambridge: Cambridge University Press.

Fudenberg, D., & Levine, D. K. (1999). *Learning and evolution in games*. Cambridge: MIT Press.
Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: gender differences. *The Quarterly Journal of Economics*, *CXVIII*, 1049–1074.
Ham, J. C., Kagel, J. H., & Lehrer, S. F. (2005). Randomization, endogeneity and laboratory experiments: the role of cash balances in private value auctions. *Journal of Econometrics*, *125*(1–2), 175–205.
Heckman, J. J. (1981). Structural analysis of discrete data with econometric applications. In *The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process* (pp. 179–195). Cambridge: MIT Press.
Hsiao, C. (2003). *Analysis of panel data*, *Cambridge books*. Cambridge: Cambridge University Press.
Ivanova-Stenzel, R., & Salmon, T. C. (2008). Revenue equivalence revisited. *Games and Economic Behavior*, *64*, 171–192.
Lehrer, S. F., Tremblay, R. E., Vitaro, F., & Schaal, B. (2004). *Raging hormones in puberty: do they influence adolescent risky behavior?* Mimeo.
Lewejohann, L., Reinhard, C., Schrewe, A., Brandewiede, J., Haemisch, A., Görtz, N., Schachner, M., & Sachser, N. (2006). Environmental bias? Effects of housing conditions, laboratory environment and experimenter on behavioral tests. *Genes, Brain and Behavior*, *5*, 64–72.
Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. *American Economic Review*, *81*(5), 1068–1095.
Sakata, S. (2002). *Quasi-maximum likelihood estimation with complex survey data*. Mimeo.
Schipper, B. C. (2011). *Sex hormones and choice under risk*. Working Paper, UC Davis.
Seinen, I., & Schram, A. (2006). Social status and group norms: indirect reciprocity in a repeated helping experiment. *European Economic Review*, *50*, 581–602.
Stern, K., & McClintock, M. K. (1998). Regulation of ovulation by human pheromones. *Nature*, *392*(12), 177–179.
Van Huyck, J., Battalio, R. C., & Beil, R. (1991). Strategic uncertainty, equilibrium selection, and coordination failure in average opinion games. *The Quarterly Journal of Economics*, 885–910.
Vanberg, C. (2008). Why do people keep their promises? An experimental test of two explanations. *Econometrica*, *76*(6), 1467–1480.
Warton, D. I., & Hudson, H. M. (2004). A MANOVA statistic is just as powerful as distance-based statistics, for multivariate abundances. *Ecology*, *85*(3), 858–874.
Williams, R. L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics*, *56*, 645–646.
Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge: MIT Press.