2006s-21

Session Effects in the Laboratory

Guillaume R. Fréchette

Série Scientifique
Scientific Series

Montréal
Novembre 2006

CIRANO
Centre interuniversitaire de recherche
en analyse des organisations

Les cahiers de la série scientifique (CS) visent à rendre accessibles des résultats de recherche effectuée au CIRANO afin de susciter échanges et commentaires. Ces cahiers sont écrits dans le style des publications scientifiques. Les idées et les opinions émises sont sous l'unique responsabilité des auteurs et ne représentent pas nécessairement les positions du CIRANO ou de ses partenaires.
This paper presents research carried out at CIRANO and aims at encouraging discussion and comment. The observations and viewpoints expressed are the sole responsibility of the authors. They do not necessarily represent positions of CIRANO or its partners.

ISSN 1198-8177

Session Effects in the Laboratory [*]

Guillaume R. Fréchette [†]

Résumé / *Abstract*

En économie expérimentale où les sujets participent à différentes sessions, les observations peuvent être plus corrélées à l'intérieur d'une même session qu'entre les participants ayant participé à différentes sessions. Ce type d'effet est présent également dans plusieurs autres domaines, autant expérimentaux que non-expérimentaux. Cette étude tente de mettre en lumière quelles sont les sources du problème et propose un ensemble de tests pour déterminer s'il y a ou non présence de corrélation intra-session dans une collecte de données, et pour identifier l'effet d'un traitement lorsque ce problème est présent. Des simulations ont été effectuées pour évaluer la performance de ces tests sur des échantillons avec peu d'observations, ce qui est commun lors de la collecte de données de type expérimental.

Mots clés : économie expérimentale, corrélation intra-session

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 problem of session effects is related to similar problems in many experimental and non-experimental fields. This paper attempts to clarify what the issues are and proposes a set of practical tests to identify the problem as well as ways to test for treatment effects in the presence of session-effects. Simulations are used to assess how these tests perform given the relatively small samples typical of experimental data sets.

Keywords: *experimental economics, session effects*

[*] I thank Martin Dufwenberg who encouraged me to write this paper. I also wish to thank Marina Agranov, Jeffrey Carpenter, Gary Charness, Yan Chen, Vincent Crawford, Pedro Dal Bo, Mark Dean, John Duffy, John Ham and Andrew Schotter for helpful comments and the C.V. Starr Center and the Center for Experimental Social Science for their financial support.

[†] CIRANO and New York University, email: guillaume.frechette@cirano.qc.ca or frechette@nyu.edu

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, siblings might exhibit more correlation than individuals across households. In experimental economics this is known as the session effects problem. To date however, there is no explicit articulation of this problem. Nonetheless, the session effects problem has become very important in experimental 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 very difficult to understand the appropriate response to it. The present paper has three aims. First, to formulate more clearly how session effects could arise and how they can be interpreted. Second, to propose ways to recognize their presence. Third, to explore the implications for experimental design and data analysis. The focus will be on practical solutions which are easy to use and can be implemented with standard software.[1]

To address the second and third goals, techniques from other fields will be used. 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 generate 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. These methods, though, are not at all the same as those used in experimental economics.

The problem of session effects is usually addressed in one of two ways in ex-

[1] Specifically, for every test and procedure discussed there exists an already programmed estimator in Stata.

perimental economics.[2] One solution is to use session averages of the variable of interest. The other solution is not to replicate the game of interest in the experiment; that is, to play the game 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 observation 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. 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 is not possible. Thus if the question of interest requires controlling for observables, this could completely rule out some questions. Also, if one is interested in the interaction of 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 (this could result in estimates which are inefficient or even not consistent). Moreover, it will be shown that both methods might create new problems.

The paper will first define the problem. Then methods for identifying and addressing the issue will be proposed. Finally a simulation exercise to illustrate and compare the performance of the different methods discussed will be provided.

Before defining the problem, three more points should be discussed. First, throughout this paper it will be assumed that the experiment uses a bewteen-subject design. Within-subject designs are not less important, and the potential for session effects is the same, but they would require a different set of solutions. Second, most tests and solutions that will be covered will be in a regression context. There are other methods for dealing with such problems, some of which will be mentioned, but focussing on regressions will make the treatment more unified. Three, examples will be used throughout, but in the interest of space, the same environment will always be used. Note, however that the observations and recommendations made in this paper are not limited to the environment of this specific example. Let us define this environment here. Subjects participate in a first-price affiliated private value auction. There are n bidders per auction and Gn subjects per session, where G is

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

the number of simultaneous auctions within a session (or the number of groups). Note that G does not have to be constant across sessions for the analysis that follows but it will be treated as such in this paper. Each subject participates in P auctions (or periods). In every auction, a value x_0 is drawn randomly from a uniform distribution with support $\underline{x}$ to $\overline{x}$. Bidders receive the private value x_{ip} where i denotes the subject and p denotes the period. x_{ip} is drawn from a uniform distribution with support $x_0 - \varepsilon$ to $x_0 + \varepsilon$. After every period, subjects are presented with the bids and values of others in the auction.[3] Groups will be denoted by g, and s denotes the session. The goal of this hypothetical experiment is to determine if changing ε from ε_0 (the control) to ε_1 (the treated) has an effect on bidding.

2 The Problem Defined

A first pass at defining session effects could be a correlation in the variable of interest across observations within a session. One cannot stop there however. In our example, if the researcher simply computed the average bid under each treatment and tested for the effect of the auction institution by a t-test, the null hypothesis of no effect of the institution could be rejected simply because the random values assigned in one treatment were higher than those in the other (this is a potential problem only because the number of sessions per treatment is usually small). Clearly, the solution here is to condition on the private values. This could be done by regression analysis or by taking the difference between the bids and values. An alternative solution in this case would be to use the same set of pseudo-random values in each treatment, in which case controlling for the treatment is equivalent to controlling for the values. This alternative, though, is not always available. Thus, the session effect problem needs to be defined as a correlation in the variable of interest (or the residual) once the relevant factors are controlled for. It could result either from some relevant factors being unobservable or from the fact that the researcher is ignorant of some relevant factors which could be controlled for if their importance was known. The greater the session effect problem, the lower is the variance in the variable of interest within a session relative to the total variance.

Two types of session effects could exist. First, imagine that less aggressive individuals shave their bids less. For example, there is evidence that testosterone in-

[3] This design was used first by Kagel, Harstad, and Levin (1987).

creases aggressiveness,[4] and that testosterone levels decrease during the day (Dabbs, 1990). Thus it could be that since different sessions are conducted at different times of day, earlier sessions display more shaving of bids, all else being equal, than those conducted later in the day. This first form of session effect will be termed a static session effect (SSE). Of course, for the purpose of this example, we assume that the researcher is not aware of this relation between testosterone levels and bidding behavior. Another example would be a difference in behavior as a function of the gender composition of a session (see for instance Gneezy, Niederle, and Rustichini, 2003). A third example, related to recent research on auction, would be that if women's bidding behavior is affected by where they are in their menstrual cycle (see Chen et al (2005) for evidence of this) and since subjects tend to be undergraduate students who often live in the same dormitories their cycles are likely to exhibit positive correlation.[5] One can also think of examples where even if the researcher is aware of such a problem, he cannot control for it because the source of the problem is not observable (for example if there is no easy or affordable way to measure testosterone levels). A concrete example of this would be that subjects might ask different questions across sessions when the experimenter is reading the instructions and those affect how all the subjects in a given session behave.

The second type of session effect emerges as a result of the interaction between subjects in a session. Imagine that subjects decide how much to shave in part by observing others' behavior during the session. If others shave more than they do, they decrease their bids, and if others shave less than they do, they increase their bids. This kind of session effect will be referred to as a dynamic session effect (DSE).

A problem with correlations which are unaccounted for is that they affect hypothesis tests. Note however that it does not affect all results equally. It is difficult to find compelling examples in experimental economics of negative within-session correlation, hence the likely problem is the rejection of the null hypothesis when it should not be rejected. Similarly, there is probably little reason to be concerned about session effects when the null hypothesis is not rejected.

As mentioned earlier, experimenters tend to deal with the session effect problem in one of two ways. One is at the level of experimental design: they do not

[4] There is direct evidence of this in rodents (Beatty, 1992). However, clear evidence of a causal relationship between testosterone and aggressiveness in humans is not as clear (see Lehrer et al, 2004 for a brief summary of the evidence).

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

repeat the game within sessions. In our example, this corresponds to using $P = 1$. Alternatively, if they deal with the problem at the data analysis level, they usually take session averages of the variable of interest and analyze (typically with a non-parametric test such as the ranksum test)[6] the data with session averages as the unit of observation. These solutions can be problematic for a few reasons.

First, if the session effect is of the static form, then not repeating the game does not resolve anything. Thus, in general, one cannot avoid solutions at the data analysis level. Second, both solutions result in a drastic loss of power. For instance, it is not unusual for P to be 10 (and often much more) and for the number of subjects to be 15, in which case, averaging by session reduces the sample by a factor of 150. If there is no session effect, then averaging by session just increases variances and leads to too few rejections of the null hypothesis. Finally, although averaging by session is usually viewed as too cautious, it may sometimes incorrectly lead to the rejection of the null hypothesis. To see this, take two cases which do not suffer from session effect problems. Suppose we want to compare value minus bids in the control and treatment. In case 1, the hypothetical experiments generate identical observations for every subject of a given session which are all lower in treatment ε_0 than in treatment ε_1. In case 2, the hypothetical experiments generate the exact same average per session as in case 1 but have extremely high variance in the observations within each session (such that there is overlap between observations in both treatments). The standard method of session averaging and using the ranksum test would produce the same test statistic for case 1 and case 2 and would reject the null of no difference if there are at least 3 sessions per treatment. It is easy to see however that one should be more confident of rejecting the null in case 1 than in case 2.

One might think that these problems could be addressed through the re-matching scheme of subjects within session. Clearly this does nothing to address the problem if it is caused by SSE, but could it help with DSE? The answer to this question would depend on what generates the DSE. DSE's could have many causes. First, it 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 which they are interacting. Second, it could result from subjects trying to influence what others do (an example of this would be strategic teaching (Camerer, Ho, and Chong, 2002). Third, it could arise because subjects (partly)

[6] Sometimes referred to as the Wilcoxon or Mann-Whitney test.

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, could also generate DSE. Clearly some of these sources of the problem, such as strategic teaching or reciprocity, can be eliminated by using a turnpike protocol (Cooper et al, 1996).[7] However, if DSE arise because of imitation, then using a fixed pairing, random rematching, a round robin procedure, or the turnpike protocol would not help. Hence, the experimenter cannot, in general, eliminate session effects through the matching protocol, although for some types of session effects it could help.

3 Tests and Solutions

3.1 Static session effects

Static session effects could take many forms, but we will allow for SSE which affects the level only not the slope coefficients. Although the latter is possible, it would complicate the analysis.[8] If the session effects only affect the level of the variable of interest, then one can estimate it in the following framework. Define y_{ip} as subject i's bid in period p, then

$$y_{ip} = \mathbf{x}_{ip}\boldsymbol{\beta} + \theta T_s + c_s + c_i + e_{ip} \tag{1}$$

where $\mathbf{x}_{ip}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{ip1} + \cdots + \beta_K x_{ipK}$ (in other words $\mathbf{x}$ is composed of K regressors and a constant term). T takes value one for the treated sessions and zero for the control.[9] c_s and c_i are respectively static session and individual unobserved effects (or SSE and IE). The usual assumption applies, but in addition we will assume that $E(c_i \mid \mathbf{x}_{ip}, T_s, c_s) = E(c_i) = 0$ and $E(c_s \mid \mathbf{x}_{ip}, T_s, c_i) = 0$.[10,11] One approach

[7] In the turnpike protocole, subjects cannot influence the decisions of future subjects they will be paired with through the decisions they take in the current match. It is sometimes referred to as a zipper design.

[8] The interested reader is referred to Ham and Kagel (2005) who also investigate techniques to deal with session-effects. Their focus is different however. They do not discuss DSE but explore more ways to deal with SSE.

[9] If there are more than two treatments, simply include one dummy variable for each treatment beside the control. Then the null hypothesis becomes the joint statistical significance of all those dummies.

[10] We will allow for IE and report some results pertaining to their estimation, but the focus will be kept on session-effects.

[11] In many cases, such as this one, the assumptions we will make are more restrictive than necessary for the estimators we will discuss to be valid.

would be to introduce a dummy variable for all but one session per treatment and to estimate equation 1 by using a random effects estimator through feasible generalized least squares (FGLS).[12] To test for the presence of session effects, one simply tests the null of joint statistical significance of the session dummies. Hence, if there are 3 sessions per treatment (session 1, 2, and 3 are in the control, 4 to 6 in the treated group), and dummy variables for sessions 1 and 4 are excluded, simply test $H_0 : c_2 = c_3 = c_5 = c_6 = 0$. If H_0 is rejected, the above allows one to determine if the treatment has an effect on the variable of interest by the statistical significance of $\frac{T+c_5+c_6}{3} - \frac{(c_2+c_3)}{3}$. In other words, one can observe whether the average level in the treated sessions is different from the control sessions. If H_0 cannot be rejected, the simpler specification without c_s can be estimated and the statistical significance of T can be established directly. This method unfortunately does not easily generalize to the case of limited dependent variables.

An alternative which does naturally extend to limited dependent variables is the following. First, assume in addition to $E\left(c_i \mid \mathbf{x}_{ip}, T_s, c_s\right) = E\left(c_i\right) = 0$ that $c_i \sim N\left(0, \sigma_i^2\right)$. Second, also assume $c_s \sim N\left(0, \sigma_s^2\right)$. This will lend itself to maximum likelihood estimation (MLE).[13] In this case, the presence of static session effects can be tested for using a likelihood ratio test.

3.2 Dynamic session effects

In recent years an important literature on peer-effects (and social networks) has developed. This literature focusses on the difficult task of disentangling selection effects from peer-effects and on ways to get correct estimates of peer-effects in the presence of the reflection problem. Researchers working on such issues also must contend with problems resulting from attrition. Dynamic session effects are essentially peer-effects, or using the terminology of Manski (1993), endogenous effects, and thus this literature will provide the basis for our testing strategy. Fortunately, in experimental data, the crux of the difficulties that appear in the peer-effects literature are either absent, or irrelevant for our purposes. First, the laboratory is exempt from selection problems since random assignment is used. Second, there is,

[12] For simplicity we assume that the time effects, if they matter, can be appropriately controlled for by including a regressor which is a function of the period.

[13] This method was used in Brandts and Cooper (2005) and Carpenter (2004) to deal with potential session effects.

for the most part, no attrition problem.[14] Third, since we will not be concerned with the estimates of session effects themselves, the reflection problem will not be an issue. Furthermore, in experiments, the researcher often knows when subjects could observe others (and what they could observe) and when they could make a choice, and those two are not simultaneous. Thus, there is no refection problem.[15,16]

To test for the presence of dynamic session effects, we estimate the following augmented version of equation 1:

$$y_{ip} = \mathbf{x}_{ip}\boldsymbol{\beta} + \theta T_s + \mathbf{z}_{g(p-1)}\boldsymbol{\delta} + c_s + c_i + e_{ip}, \qquad p > 1 \tag{2}$$
$$\text{and } E\left(c_i \mid \mathbf{x}_{ip}, T_{ss}, \mathbf{z}_{g(p-1)}, c_s\right) = E\left(c_i\right) = 0; \quad E\left(c_s \mid \mathbf{x}_{ip}, T_s, \mathbf{z}_{g(p-1)}, c_i\right) = 0$$

where $\mathbf{z}_{gp}$ are constant at the group level in a given period. In an experiment, since subjects do not (usually) interact within a period before making a decision, the relevant variables are those from the previous period. These could include, for instance, group averages of $\mathbf{x}$ in the previous period and group averages of y in the first period. To test for the presence of session effects, we would look for the statistical significance of elements of $\boldsymbol{\delta}$.

Again the c_s can be controlled for as a set of indicator variables in a FGLS framework or as a random-effect in a MLE framework. If MLE is used, then we will want to allow for a more flexible form than what has been specified previously. Namely $c_s \sim N\left(\mathbf{z}_{sg}\boldsymbol{\psi}, \sigma_s^2\right)$ where $\mathbf{z}_{sg}$ now also includes variables which are constant at the group level only (a correlated random effects estimator, see Chamberlain 1982, 1984).

If we confirm the presence of DSE and if the researcher is confident about the exact form that these interactions take, they can be appropriately controlled for. A more modest approach is not to directly control for the DSE but rather adjust the variance-covariance matrix of the estimates to account for more general forms of correlation in the error terms. Although this is less efficient under the correct specification, it is much more robust to misspecifications. Thus, once the presence of potentially complicated DSE is confirmed – for which the experimenter usually

[14] Exceptions to this include experiments with bankrupcies and experiments which use experienced subjects.

[15] This is used by Fortin, Lacroix, and Villeval (2004) to study tax evasion and social interactions in the laboratory.

[16] This is not to say that all problems are eliminated, only that many of the hurdles encountered in this literature using field data are absent here.

does not have a model – it might be sensible to use a less efficient estimation method which is robust to many correlation structures of the error term. Remember that we assume the variable of interest is T. If one is interested, for instance, in belief formation, then the approach to be outlined below is probably insufficient.

The estimation equation would now be simplified to

$$y_{ip} = \mathbf{x}_{ip}\boldsymbol{\beta} + \theta T_s + e_{ip}, \tag{3}$$

but the (session clustered) variance of the estimator is now estimated by

$$\widehat{V}_s = a\widehat{V}\left(\sum_{k=1}^{S} u_k' u_k\right)\widehat{V}$$

where a is a finite sample adjustment, $\widehat{V}$ denotes the conventional estimator of variance and u_k is the contribution of the k^{th} session to the vector of scores.[17] This allows for arbitrary correlations within sessions.[18] In other words, if the error terms are stacked by session, the variance-covariance of the error term is assumed to be a block diagonal matrix. One could alternatively assume a more specific structure, for instance a fixed correlation for a given subject within a session and a different correlation (but the same across periods and subjects) across subjects of a same session. This would be more efficient if the assumed structure is correct but less robust. As before, the statistical significance of θ is determined using a Wald test.

4 Comparison

We will compare the tests and estimators presented in this paper to give an impression of their performance. This will be performed in the context of our example. There is no pretence of generalizability of the results of these simulations, but hopefully they illustrate the trade-offs between different ways of assessing treatment effects. It also will give an idea of the performance of estimators which rely on asymptotic arguments in small samples typical of experiments.[19]

The fictional experiment will have 4 groups of 4 subjects per session, 4 sessions

[17] For the specific formulas used, refer to StataCorp (2003).

[18] Such a correction is performed for instance in Chen et al (1995).

[19] No attempt will be made to determine the "optimal" number of repetitions, size of groups, etc. as this would depend too much on the specific experiment (game, correlations in the data...).

per treatment, and 30 periods per session.[20] Subjects are randomly re-matched across periods. As in Kagel, Harstad, and Levin (1987): $\underline{x} = 25$ and $\overline{x} = 125$, and ε will be either 12 (the control) or 24 (the treated).[21]

Different types of bidders will be simulated. The no IE and no session effects types use the bid function

$$\begin{aligned} b_{ip}(x_{ip}) &= x_{ip} - (6 + 12\theta T_s) + e_{ip} \\ e_{ip} &\sim N(0,1) \end{aligned} \tag{4}$$

which is almost the risk-neutral symmetric Nash equilibrium bid function if $\theta = 1$ and $e_{ip} = 0$.[22] Although this specification implies sometimes bidding above value (which is clearly irrational), this occurs with such small probability (about 1 in 1014713328 for the control) that it should not be problematic. In the interest of simplicity this ignores other factors which have been shown to be relevant in reality such as the effect of time and cash-balances (Ham, Kagel, and Lehrer, 2004). IE and SSE will be included in the following bid function

$$\begin{aligned} b_{ip}(x_{ip}) &= x_{ip} - (6 + 12\theta T_s) + c_s + c_i + e_{ip} \\ e_{ip} &\sim N(0,1), c_i \sim N(0,0.4), c_s \sim N(0,0.2). \end{aligned} \tag{5}$$

The data generating process (DGP) for simulated subjects with IE, SSE, and DSE is given by

$$b_{ip}(x_{ip}) = x_{ip} - \frac{3(6 + 12\theta T_s)}{4} - \frac{1}{4}\frac{\sum_{j \in g, j \neq i}\left(x_{j(p-1)} - b_{j(p-1)}\right)}{3} + c_s + c_i + e_{ip}, \tag{6}$$

$$\text{for } p > 1,$$

$$e_{ip} \sim N(0,1), c_i \sim N(0,0.4), c_s \sim N(0,0.2),$$

and bidding is given by 5 in period 1. That is to say after period 1, subjects' bids are also influenced by the average amount others in their group shaved in the

[20] Ham, Kagel, and Lehrer (2004) use groups of 4 or 6 subjects for 30 periods in their affiliated private value auction experiment. They, however, have 2 groups per experiment. Sessions of 16 subjects was used simply because it was thought to be more representative of the modal experiment.

4 sessions per treatment is quite common. Certainly, few experiments go below 3 as tests such as the ranksum test require at least 6 observations to reject an hypothesis at the 5% level. Thus, if one uses session averages, 3 sessions per treatment is the minimum.

[21] Kagel, Harstad, and Levin (1987) also had $\varepsilon = 6$.

[22] To be the equilibrium bidding function it would need to add $\frac{\varepsilon}{10}e^{-\left(\frac{2}{\varepsilon}\right)\left[x_{ip} - (\underline{x} + \varepsilon)\right]}$ – but that quantity quickly becomes negligible – and it only applies to $\underline{x} + \varepsilon \leq x_{ip} \leq \overline{x} - \varepsilon$.

Estimation	Simulated Data			Correct Detection at 5%		
Method	IE	SSE	DSE	IE Test	SSE Test	DSE Test
FGLS with Individual RE, SSE dummies and allowing for DSE	no	no	no	93	95	94
	yes	no	no	100	91	96
	yes	yes	no	100	97	96
	yes	yes	yes	100	97	100
	no	yes	no	93	100	94
	no	yes	yes	93	100	100
	no	no	yes	93	95	100
	yes	no	yes	100	91	100

IE, SSE, and DSE stand for individual, static session and dynamic session effect respectively.

All numbers are percentages.

The numbers are the same for $T = 0$ and $T = 1$.

Table 1: Percentages of Time Each Test Correctly Infers the Presence of IE, SSE, and DSE

previous period.[23] Every other simulated type can be obtained by removing c_i or c_s in 5 or 6.

For every type, the case of $\theta = 1$ and $\theta = 0$ will be considered. That is the case where there is a treatment effect and the case where there is not. 1000 replications are performed in each simulation.

The tests for the presence of session effects (and IE) will be performed using estimates from FGLS allowing for IE, SSE, and DSE. Specifically, we will estimate $b_{ip} = \alpha + \mathbf{x}_{ip}\boldsymbol{\beta} + \theta T_s + \delta \frac{\sum_{j \in g, j \neq i} \left(x_{j(p-1)} - b_{j(p-1)}\right)}{3} + c_2 + c_3 + c_5 + c_6 + c_i + e_{ip}$ where c_i is a random effect and period 1 is dropped. The tests will be Wald tests except the test for the statistical significance of the IE which is the Breush-Pagan Lagrange multiplier test.[24]

Table 1 reports the percentage of time each test correctly identifies the presence or absence of IE, SSE, or DSE for the most general estimation equation we will consider. All three tests clearly perform well despite the small samples, correctly identifying the presence or absence of IE, SSE, and DSE an average of 96.5%, 95.75%,

[23] The values and bids of others in their group is part of the feedback received in standard auction experiments.

[24] FGLS estimates are obtained using the xtreg command (StataCorp, 2003).

			Estimation Method							
Simulated Data			FGLS with Individual RE, SSE dummies and allowing for DSE		FGLS with Individual RE and SSE dummies		FGLS with Individual RE and allowing for DSE		FGLS with Individual RE	
IE	SSE	DSE	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
no	no	no	4	0	4	0	4	0	4	0
yes	no	no	7	0	6	0	6	0	6	0
yes	yes	no	52	0	50	0	46	0	43	0
yes	yes	yes	52	48	61	28	46	53	53	36
no	yes	no	81	0	81	0	72	0	65	0
no	yes	yes	81	19	86	12	69	25	67	25
no	no	yes	4	96	14	0	4	96	13	0
yes	no	yes	8	92	17	27	6	94	16	28

IE, SSE, and DSE stand for individual, static session, and dynamic session effect respectively.
All numbers are percentages.

Table 2: Error Frequency at 5 Percent Level (When Testing for a Treatment Effect)

and 97.5% of the time respectively. Of course the specific percentages are a function of the parameters in the simulation and for a different simulation they could perform better or worse.

We must now consider (1) how effective the different estimators presented are at identifying the presence or absence of a treatment effect in the presence of session effects and (2) is their performance worse than that of the usual method. To answer this second question we will compare the estimator described to the performance of the ranksum test on session averages of the difference between bid and value.[25]

The results are reported in Tables 2 and 3. They report the percentages of Type I error, rejecting the null of no treatment effect when $T = 0$, and of Type II error, not rejecting the null of no treatment effect when $T = 1$. Results for ordinary least squares (OLS) with individual clusters and OLS are reported for comparison. Note that depending on the estimator and the DGP, poor performance can be the result of (a) misspecification or (b) small sample.

The first observation we make is that SSE are very problematic in that, except for OLS with session clusters and the ranksum test on session averages, the presence

[25] Other tests are sometimes used. For instance Feltovich (2003) suggests using the robust rank order test, but the ranksum test is the most commonly used.

			Estimation Method							
Simulated Data			OLS with Session Cluster		OLS with Individual Cluster		OLS		Ranksum Test	
IE	SSE	DSE	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
no	no	no	7	0	5	0	4	0	6	0
yes	no	no	8	0	6	0	48	0	8	0
yes	yes	no	9	0	43	0	78	0	7	0
yes	yes	yes	8	86	52	37	83	14	7	89
no	yes	no	8	0	64	0	80	0	6	0
no	yes	yes	8	85	66	25	85	13	6	87
no	no	yes	7	0	15	0	14	0	6	0
yes	no	yes	9	49	15	29	60	9	8	55

IE, SSE, and DSE stand for individual, static session, and dynamic session effect respectively.
All numbers are percentages.

Table 3: Error Frequency at 5 Percent Level (When Testing for a Treatment Effect)

of SSE leads to a very high rate of Type I error. This is clearly a result of the extremely small samples (of sessions per treatment) involved. In the case of OLS with session cluster and the ranksum test, lower Type I error rates come at the cost of very high Type II error rates if the SSE coexist with DSE but are otherwise not problematic.

DSE on their own result in low (slightly higher but almost at the 5% level of the test) levels of Type I error for OLS with session cluster and the ranksum test. These can be reduced marginally by using FGLS and allowing for DSE, but this comes at the cost of important increases in Type II error rates.

In addition, we also report the performance of using a conditional approach. Namely, estimate FGLS allowing for IE, SSE, and DSE, and test for the presence of each. If either the test for the presence of SSE or DSE rejects the null, then use estimates from OLS with session cluster. If the null of no IE is rejected, use estimates from FGLS (without session dummies or regressors for DSE), and otherwise use OLS estimates. These are reported in Table 4. Overall, the performance is similar to that of OLS with session clusters or the ranksum test on session averages.

These results suggest no clear benefit of using session averages as opposed to OLS with clustered standard errors or to using the conditional procedure described in the preceding paragraph. Note that this is true even though the sample size are

Estimation	Simulated Data			Error Freq. at 5%	
Method	IE	SSE	DSE	Type I	Type II
Conditional	no	no	no	5	0
Procedure:	yes	no	no	5	0
FGLS w/ IE,	yes	yes	no	9	0
SSE, and DSE	yes	yes	yes	8	86
first, then either	no	yes	no	8	0
OLS cluster,	no	yes	yes	8	85
FGLS w/ IE	no	no	yes	7	0
or OLS	yes	no	yes	9	49

IE, SSE, and DSE stand for individual, static session, and dynamic session effect respectively.

All numbers are percentages.

Table 4: Error Frequency at 5 Percent Level (When Testing for a Treatment Effect)

relatively small.

5 Conclusion

Clearly one cannot generalize on the basis of the specific numbers obtained in this particular simulation. Nonetheless they do suggest that OLS with session clusters performs similarly to the ranksum test on session averages. Since the former is much more flexible than the latter, and using session averages loses information which may create problems (something that does not happen with OLS with session clusters), these simulations do seem to validate hypothesis testing that does not rely on session averages only.[26]

As discussed earlier, one standard method for dealing with SE is to play the game of interest only once, but this does not address SSE and eliminates experiments where experience is important. Nonetheless, it is interesting to see how the performance of this approach would compare. To do this in our simulations, we test for the presence of a treatment effect using the ranksum test on period one data only. This keeps the number of sessions and subjects constant (not the cost). Clearly, one could increase the number of sessions to counterbalance the drop in the number of

[26] Note that standard sofware (such as Stata) also allows for t-tests and ANOVA estimation with corrections equivalent to the use of session clustering in a regression context.

observation. Nonetheless the parameters used in this experiment are not unusual even if there is only one period per session. For the particular simulated experiments considered in this paper, Type I error would be between 33% and 34% and Type II error about 61% for every DGP. Type I error is increased to 34% (from 33%) when SSE effects are present. Clearly, SSE did not dramatically decrease the performance despite the fact that we use only period 1 observations. This being said, overall performance is drastically diminished (as compared to OLS with session clusters).

The simulations suggest that SSE might be extremely difficult to deal with (given the standard experimental sample size). However, of the two types of session effects we have discussed, they do seem to be the least likely to occur. As mentioned earlier SSE could also take different forms (such as affecting slope parameters). If these different forms are really a concern, then session clustering can be used.

As for DSE, although we have described a method for testing for their presence and accounting for them, we have sidestepped the question of misspecification. Implicitly we have assumed that the experimenter knows the structure of the DSE. In some cases this is plausible. For instance, in a minimum game (Huyck, Battalio and Beil, 1990), if the subjects' only received feedback is the minimum of the group, it is sensible to think that any DSE could only work through the minimum of the group. However, there are experiments where subjects receive extremely detailed feedback about the groups behavior, making it more difficult for the experimenter to know what, if anything, is relevant. In such cases where the experimenter has reasons to believe there are DSE, but little prior idea about the form they might take, simply correcting the variance is probably advisable.

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 DSE (by eliminating feedback) and allow the best chance to identify SSE (through the many repetitions and with the numerous sessions). Unfortunately, eliminating feedback could very well hinder learning (see for instance Armantier (2004)).

In our view, the results of the simulations suggest that there is no need to restrict analysis by imposing the use of session averages. This is not to say that session effects should be ignored, but rather that there are tools that allow such correlations but are not as restrictive.

References

Altman, D. G., and J. M. Bland (1997): "Statistical Notes: Units of Analysis," *BMJ*, 314, 1874.

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 *Handbook of Behavioral Neurology volume 2*, ed. by Arnold A. Gerall, Howard Moltz and Ingeborg L. Ward. New York: Plenum Press.

Brandts, J., and D. J. Cooper (2005): "A Change Would Do You Good... An Experimental Study on How to Overcome Coordination Failure in Organizations," mimeo.

Camerer, C. F., T.-H. Ho, and J.-K. Chong (2002): "Sophisticated EWA Learning and Strategic Teaching in Repeated Games," *Journal of Economic Theory*, 104, 137–188.

Carpenter, J. P. (2004): "The Demand for Punishment," mimeo.

Chamberlain, G. (1982): "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, 18, 5–46.

——— (1984): "Panel Data" in *Handbook of Econometrics, Volume II*, ed. by Z. Griliches and M. D. Intriligator. Elsevier Science Publishers, 1247-1318.

Chen, Y., P. Katuscak, and E. Ozdenoren (2005): "Why Can't a Woman Bid More Like a Man," mimeo.

Cooper, R., D. V. DeJong, R. Forsythe, and T. W. Ross (1996): "Cooperation without Reputation: Experimental Evidence from Prisoner's Dilemma Games," *Games and Economic Behavior*, 12, 187–218.

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

Davis, D. D., and C. A. Holt (1993): *Experimental Economics*. Princeton University Press, Princeton, NJ.

Excoffier, L., P. E. Smouse, and J. M. Quattro (1992): "Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data," *Genetics*, 131, 479–491.

Feltovich, N. (2003): "Nonparametric Tests of Differences in Medians: Comparison of the Wilcoxon-Mann-Whitney and Robust Rank-Order Tests," *Experimental Economics*, 6, 273–297.

Fortin, B., G. Lacroix, and M.-C. Villeval (2004): "Tax Evasion and Social Interactions," IZA working paper.

Friedman, D., and S. Sunder (1994): *Experimental Methods: A Primer for Economists*. Cambridge University Press, Cambridge.

Gneezy, U., M. Niederle, and A. Rustichini (2003): "Performance in Competitive Environments: Gender Differences," *Quarterly Journal of Economics*, CXVIII, 1049–1074.

Ham, J. C., J. H. Kagel, and S. F. Lehrer (2005): "Randomization, Endogeneity and Laboratory Experiments: The Role of Cash Balances in Private Value Auctions," *Journal of Econometrics*, 125(1-2), 175–205.

Huyck, J. V., R. C. Battalio, and R. Beil (1990): "Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure," *American Economic Review*, pp. 234–248.

Kagel, J. H., R. M. Harstad, and D. Levin (1987): "Information Impact and Allocation Rules in Auctions with Affiliated Private Values: A Laboratory Study," *Econometrica*, 55(6), 1275–1304.

Lehrer, S. F., R. E. Tremblay, F. Vitaro, and B. Schaal (2004): "Raging Hormones in Puberty: Do They Influence Adolescent Risky Behavior?," mimeo.

Manski, C. F. (1993): "Identification of Endoegous Social Effects: The Reflection Problem," *The Review of Economic Studies*, 60(3), 531–542.

Sakata, S. (2002): "Quasi-Maximum Likelihood Estimation with Complex Survey Data," mimeo.

StataCorp (2003): *Stata Statistical Software: Release 8.0*chap. 23.14 Obtaining Robust Variance Estimates, pp. 270–275. Stata Corporation, College Station, TX.

Stern, K., and M. K. McClintock (1998): "Regulation of Ovulation by Human Pheromones," *Nature*, 392(12), 177–179.

Warton, D. I., and H. M. Hudson (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.