Clustering standard errors at the "session" level*

Duk Gyoo Kim†

July 17, 2021

Abstract

Session-specific features of a laboratory experiment, if those exist, do not disappear by clustering standard errors at the session level. Randomly ordering sessions to deal with sampling issues, cannot justify clustering the standard errors at the session level. The experimental design, reflected on the researchers' intention, should primarily determine the clustering level. In a typical controlled laboratory experiment where subjects make choices in the same environment repeatedly, clustering at a participant level is inherited from the experimental design, and standard errors could be larger (that is, a statistical inference could be more conservative) when clustered at the individual or decision-group level than the session level. It implies that clustering standard errors at the session level can lead to false-positive treatment effects if it is mistakenly chosen. Having a small per-session sample to increase the number of sessions could yield undesirable heterogeneities that are hard for the experimenter to control or observe.

Keywords: Lab experiment, Cluster-robust standard errors, Statistical inference

JEL codes: C18, C90

1 Introduction

This paper aims to convince the experimental economists and the readers interested in lab-experimental studies that the session-level clusters should be used only in particular situations with proper justification. A session is typically defined as a group of individuals who participate in the same laboratory experiment simultaneously. For an experiment adopting a between-subject design, a subject participated in one session

*I thank Guillaume Fréchette, Franziska Heinicke, Sang-Hyun Kim, Wooyoung Lim, Yoshiyasu Rai, Euncheol Shin, Donggyu Sul, Wladislaw Mill, and the participants at Mannheim/ZEW Experimental seminar, the Korean Economic Review International Conference, and 2020 ESA Global Online Around-the-Clock Conference for their helpful comments, and Elisa Casarin for her research assistance.

†Department of Economics, University of Mannheim, *d.kim@uni-mannheim.de*

only,[1] so a set of observations from an individual is a proper subset of the entire sample from a session, which is a proper subset of the entire sample from the same treatment. Thus, adding individual- or session-fixed effects on the regression does not help us examine a treatment effect due to multicollinearity.

Obtaining accurate standard errors of the treatment effect is fundamental for proper statistical inference. Although many studies discuss the proper use of cluster-robust standard errors (e.g., Cameron et al., 2008; Abadie et al., 2017; de Chaisemartin and Ramirez-Cuellar, 2020), to the best of my knowledge, only a few studies, including Moffatt (2016) explicitly discuss it within the context of laboratory experiments.[2] Perhaps it is why we see some researchers report standard errors clustered at the session level and some at the individual level. Among all 322 published papers using lab-experimental data at the *Experimental Economics* from March 2010 to March 2020, 124 papers mentioned cluster-robust standard errors. Standard errors of 40 papers are clustered at the participant level, and those of 34 papers are at the session level.[3]

It is often argued that standard errors should be clustered at the session level concerning the session-specific effects. For example, Keith Marzilli Ericson, a co-editor of the *Journal of Public Economics*, points out that many lab-experimental papers fail to randomly assign participants to treatment, with claiming that once researchers "[d]o session-level randomization,"[4] then the statistical "[i]nference should cluster standard errors at the session level."[5] Also, it is not uncommon that reports from referees point out that the standard errors should be clustered at the session level. Most of the time, their reasoning, including ones that Ericson made on his blog post, is that there might be some "static" session effects (Fréchette, 2012).[6] This reasoning—using session-level cluster adjustment for session effects—is not on solid ground. A concern for static session effects is the reason for randomizing or counterbalancing the sessions so that the session-specific idiosyncratic features can be integrated out, not for clustering standard errors at the session level.[7] I am worried that many researchers seem to use session-

[1] On the contrary, a within-subject design assigns a participant to two or more treatments. In this case, considering session-level clusters is even less persuasive as the design's primary purpose is to examine individual changes.

[2] Moffatt (2016) explains that researchers can consider different (subject-level as the lowest and session-level as the highest) clustering. When analyzing example data, he uses subject-level clustering only.

[3] Some papers use exogenously given clusters, such as classes and cohorts. Other papers used cluster-adjusted standard errors when analyzing empirical data, not experimental data. A few papers consider a fixed independent group as a clustering unit, which I will discuss in Section 3.

[4] In a typical setting, one session is conducted at one time, so session-level randomization practically implies randomly ordering the sessions.

[5] In his blog post (Ericson, 2018, link) are more details.

[6] The static session effects can be understood as a realization from a noise distribution that affects the observational outcomes in level. The dynamic session effects can be understood as the observational relationships across subjects within a session due to the feedback from interactions with other participants.

[7] In the context of randomized field experiments, de Chaisemartin and Ramirez-Cuellar (2020) similarly

level clustered standard errors to remedy session effects, without further justifying why a session should be the cluster level. The experimental design should determine the clustering level,[8] and only when the design inherits the positive observational relationship within a session, standard errors should be clustered at the session level. Figure 1 summarizes my arguments.

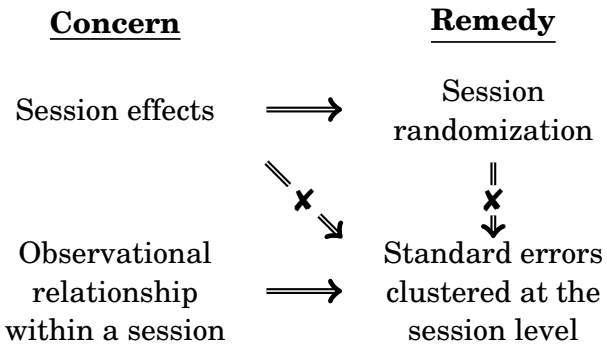


Figure 1: Clustering at the session level is not a remedy for session effects.

By inheritance of the experimental design, I mean what the researchers *intended* at the experimental design stage. When the experiment adopts random re-matching, the researchers intend to disconnect (or minimize) the dynamic relationships between decision rounds. If the session size is sufficiently large, the random rematch will approximate perfect stranger matching. Even when the session size is small, the random rematch prevents subjects from considering dynamic strategies because they do not know who were the previously matched players in what decision round. When the researchers allow the subjects to play the same game repeatedly, unless the researchers expect the subjects to play in a completely random manner, they intend the positive observational dependence within a subject. However, the researchers do not *design* the direction of the possible dynamic session effects. Of course, the researchers may expect some interactions among subjects will affect their decisions, but the directions of such interactions are not designed. Take the public goods game with a perfect stranger match, for example. Suppose one subject observes more contributions from others in one decision round. Would he increase his contribution later because he wants to be a conditional cooperator, or decrease it because he observes that the public goods are well provided without his contributions and the free-riding incentives become salient? When the experimental design allows subjects to play the public goods game repeatedly, is the direction of the interactions designed as well? If answers to both questions are negative, then the

claim that clustering standard errors at the unit-of-randomization level may lead to a severe downward bias of the variance estimator of the treatment effect.

[8] My claim is in line with Abadie et al. (2020), who claim to consider design-based uncertainty instead of sampling-based one for statistical inference.

observational relationship within a session is not inherited from the experimental design. Moreover, as I will elaborate later in Section 5, the negative relationship between a subject's decision and the decisions of the previous group members would substantially *exacerbate* the type-1 error.

To minimize destructive discussions, I emphasize two things that I am *not* claiming. First, I am not claiming that we should not worry about static session effects. The experimenter's crucial responsibility is to maintain every session's environment as homogeneous as possible, except for the treatment conditions being examined. Since it is challenging, if not impossible, to make every session environment identical, the experimenter must make sure both the control-group participants and the treatment-group participants are from the same population by randomizing or counterbalancing the session orders. In this regard, I entirely agree with what Ericson wrote in his blog: "Your subject population could be changing over time (perhaps early subjects are more eager, or have lower value of time). Or news events could change beliefs and preferences. The list of potential stories can be long; some can be ruled out, others cannot." Indeed, the potential stories are long: Perhaps one experimenter manages sessions better than another experimenter. Subjects participating in an early morning session may have distinctive characteristics than other subjects. An exogenous aggregate shock (e.g., COVID-19 pandemic) may arise between sessions. Some sessions may be conducted in more disturbing situations due to unexpected constructions, delays caused by technical glitches, or unexpectedly high/low temperatures, to name a few. Thus, it is legitimate for readers, editors, and referees to demand more sessions if they are concerned about potential static session effects. For similar reasons, a sequential modification of the experimental design—earlier sessions conducting X and Y and (perhaps several months) later sessions conducting X' and Z—may significantly undermine the internal validity of the research. Although I am wholly sympathetic to the concerns about the static session effects, it is a reason for being careful about sampling subjects from the same population pool by randomizing the sessions, a reason for making session environments as homogeneous as possible, and a reason for checking and controlling for session-particular features, but not the reason for clustering standard errors at the session level.

Second, I am not claiming that clustering standard errors at the session level is futile, especially when the experiment exploits *positive* interactions among subjects in a session. A session-level cluster can undoubtedly address the "dynamic" session effects or the observational dependence within the session. It is sometimes tightly aligned with the experimental design, especially when the subjects made decisions only once or the session-(or "market"-)level interactions are of the main interest.[9] Although Fréchette

[9]For example, Engelmann and Hollard (2010) have participants who made only a small number of deci-

(2012) argues for using standard errors clustered at the session level when there is "only one observation per subject so that we do not need to keep track of the periods" (p. 488),[10] it should not be merely extrapolated to a case where there are many observations per subject. Thus, this paper can be understood as an extension of his paper. Again, the current paper focuses less on the studies where the experimental design inherits strong and positive dynamic interactions within a session, which I believe the session-level clustering is appropriate, but it focuses on the discussions about the proper cluster level when individuals in the lab make repeated decisions.

The rest of this paper is organized as follows. Sections 2 and 3, without formal expositions, illustrate why standard errors need to be clustered and why clustering at the session level should be considered carefully. Section 4 presents a simple econometric model to pinpoint the issues in choosing proper cluster levels. Section 5 shows some Monte-Carlo simulation results. Section 6 discusses practical issues regarding cluster-robust standard errors for the laboratory data, and Section 7 concludes.

2 Why do we cluster standard errors?

Clustered standard errors should be considered when observations within a cluster are related to each other. In other words, if the observations within a cluster are similar, then the errors within a cluster will be more correlated than those of the entire sample. Thus, without "penalizing" the observational similarity, we will have downward-biased standard errors, leading to false-positive treatment effects more often. Throughout the entire paper, I consider situations with no true treatment effects. Thus, by a false positive, I mean an error to mistakenly reject the true null hypothesis (no treatment effect).

To elaborate on what I mean by "penalizing" similarity, consider the following. There are ten observations: five from a control group experiment, and the other five from the treatment group experiment. Assume that except for the treatment condition, everything is homogeneous and appropriately controlled. A researcher tests if the mean control-group observation is different from the mean treatment-group observation.

sions and focus more on the interaction within a session. Cipriani et al. (2017)'s interest is on the session-level information contagion, so the interactions within a session are inherited from the design. Corgnet et al. (2018) similarly justify their use of session-level clustering because each experimental market features a zero-sum game where an increase in one trader's earnings mechanically reduces other traders' possible gains within a session. Bracha et al. (2015) and Carpenter (2016) experimentally examine the attributes of labor supply, which is the accumulation of an individual's decisions, so it is pertinent to regard the labor supply as one observation per subject.

[10]This restriction is judicious because Fréchette (2012) focuses on the discussions about the session effects, not the relative importance of subject-specific effects and the session effects.

	Control session					Treatment session				
ID	1	2	3	4	5	6	7	8	9	10
Obs.	2	1.2	1.5	1.7	2.5	2.6	1.8	1.5	1.3	2.1

A standard t-test does not reject the null hypothesis that two means are the same (p-value=0.8085). The standard error of the mean difference is 0.3194.

Now, suppose that the researcher's half-sleeping RA mistakenly duplicated the observations several times.

	Control session					Treatment session				
ID	1	2	3	4	5	6	7	8	9	10
Obs.	2	1.2	1.5	1.7	2.5	2.6	1.8	1.5	1.3	2.1
Obs.2	2	1.2	1.5	1.7	2.5	2.6	1.8	1.5	1.3	2.1
Obs.3	2	1.2	1.5	1.7	2.5	2.6	1.8	1.5	1.3	2.1
					⋮					
Obs.50	2	1.2	1.5	1.7	2.5	2.6	1.8	1.5	1.3	2.1

The (un-clustered) standard error of the mean difference is 0.040, and the null hypothesis is rejected (p=0.0487). This inference is obviously wrong because it ignores the perfect correlation between observations at the participant level. The standard error clustered at the participant level is 0.3014, and the treatment effect becomes insignificant again (see Table 1).

Obs	(0)	(1)	(2)
Treatment	0.0800	0.0800**	0.0800
	(0.25)	(1.98)	(0.27)
_cons	1.780***	1.780***	1.780***
	(7.88)	(62.19)	(8.49)
Cluster SE	–	–	ID
N	10	500	500

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1: A false-positive effect when SEs are unclustered.

Although the example above is too unrealistic because of the perfect correlation between observations within a cluster, we can draw one clear takeaway message. A researcher must consider clustering standard errors when observations within a cluster are expected to be related. That is the way of providing more robust statistical results.

A naturally followed question is what the proper cluster level would be. Unlike other empirical studies where the clustering units can be multi-dimensional, potential clusters in a between-subject experiment—individual or session—are uni-dimensional: The set of individual-level observations is a proper subset of the set of session-level observations. In the following sections, I claim that if the lab experiment asks the participants to make decisions in a similar environment repeatedly, clustering at the participant level is inherited from the experimental design, so it is unnatural to cluster standard errors at the session level.

3 Illustration: Is the session-level clustering robust?

If standard errors clustered at the session level are larger than those at the individual level, it means that the session-level observations are more correlated than the individual's repeated choices. It may not be the case when the subjects are asked to make decisions in the same environment repeatedly.

To support my claims, I use hypothetical data. Although I found some studies reporting standard errors clustered at the session level, and sometimes the statistical results become less significant when clustering the standard errors differently, I do not intend the current paper to be read as comments/criticism to those papers. It is the main reason for using hypothetical data. However, all hypothetical data aim to be plausible and capture the key features from actual data.

Imagine a particular type of controlled lab experiment on a group decision making,[11] where a between-subject design, random rematch, anonymity, and no communications are adopted. To be more illustrative, suppose that six subjects per session have ten repeated decision rounds choosing an integer between 1 and 50, and the payoff of each round is determined by the subject's decision, a randomly-matched pair's decision, and some luck. At the beginning of a new round, the subjects are randomly rematched with another subject in the session. Their decisions are made anonymously, and they are not allowed to communicate with each other. Each subject participates in only one session. Suppose a researcher collected data from four (two control and two treatment) sessions,[12] as shown in Table 2.

Each column of Table 2 is a vector of an individual's decisions over ten rounds. A

[11] For an experiment where a single player makes a streak of decisions under some uncertainties, it is straightforward to cluster standard errors at the individual level. Here I focus on experiments involving strategic decisions.

[12] Admittedly, the example here contains too few samples (six subjects in each session of four). The mere purpose is to display the entire observations in a small table, and I do not intend to use the small sample properties. See the Monte-Carlo simulation results.

	Control-Session-01						Treatment-Session-01						Control-Session-02						Treatment-Session-02					
Round\ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
r01	6	36	21	17	32	17	29	22	13	40	19	29	37	7	16	23	26	16	8	41	29	16	33	16
r02	12	42	23	20	27	15	27	16	15	39	13	23	39	8	19	23	31	11	11	39	24	20	29	14
r03	6	36	28	18	31	18	29	16	14	43	17	31	39	5	13	23	29	11	13	39	30	22	34	14
r04	12	37	26	14	31	17	27	18	13	42	19	30	35	6	12	26	25	14	9	38	25	16	31	15
r05	6	43	22	19	27	17	34	17	10	42	16	25	42	8	16	21	30	12	8	44	22	20	28	20
r06	7	40	22	15	29	12	31	17	9	36	16	30	40	8	16	24	27	15	9	42	25	17	32	18
r07	9	36	22	20	32	17	27	19	13	38	14	29	39	11	18	23	27	14	14	38	28	17	34	17
r08	9	37	23	19	33	15	32	21	12	41	13	23	37	6	14	25	29	11	9	44	25	20	33	20
r09	11	39	21	17	29	16	35	17	11	39	19	27	38	5	14	20	26	14	14	40	24	22	32	18
r10	10	41	20	16	27	16	31	19	10	43	19	28	41	12	12	18	27	16	15	41	27	20	31	17
Std.	2.4	2.7	2.4	2.1	2.3	1.7	2.9	2.0	1.9	2.3	2.5	2.9	2.1	2.4	2.4	2.4	1.9	2.0	2.7	2.2	2.5	2.3	2.0	2.2
	Std.(session)=11.2						Std.(session)=11.4						Std.(session)=11.4						Std.(session)=11.6					
Std.(whole)=11.4																								

Table 2: Data from Four Sessions

researcher wants to examine the mean treatment effect. If we do not cluster standard errors, the mean control-group observation is significantly different from the mean treatment-group observation ($\bar{y}_C$=21.55, $\bar{y}_T$=24.15, t=1.9808, p-value=0.0488). The standard error of the difference is 1.313.

In the hypothetical data, standard deviations of the individual-level observations are small, which implies that they made similar choices over the rounds. The session-level standard deviations are as large as the standard deviation of the entire sample. If we cluster the standard error of the mean difference at the participant level, the difference is no longer statistically significant (models (2) and (5) in Table 3). However, clustering standard errors at the session level does not handle the false-positive treatment effect (models (3) and (6) in Table 3.)

	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	2.600**	2.600	2.600***	2.600	2.600	2.600***
	(1.98)	(0.63)	(4.43)	(0.62)	(0.63)	(4.43)
_cons	21.55***	21.55***	21.55***	21.55***	21.55***	21.55***
	(23.22)	(7.26)	(36.75)	(7.23)	(7.26)	(36.75)
Individual RE	No	No	No	Yes	Yes	Yes
Cluster SE	–	ID	Session	–	ID	Session
N	240	240	240	240	240	240

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3: False-positive effects when SEs are clustered at a session level

Unless the experiment encourages every subject to make completely arbitrary decisions, the observational similarity at the participant level is *inherited from* the experimental design when the experiment asks a participant to make repeated decisions. Two

of the primary reasons for the repetitions are to increase the number of observations and to allow subjects to learn the equilibrium of the game in the course of getting feedback. Thus, when the learning effects are not of their primary interest, researchers often focus on the observations from the latter half decision rounds. Those observations are likely "less noisy," meaning that the individual's decisions are similar over rounds. Roughly put, the observations become similar to the half-sleeping RA's duplicated data.

Instead of a random rematch, if the experiment involves fixed independent groups of the participants over the repeated decision rounds, then clustering standard errors at the group level could also be considered. If the experiment features repeated games (e.g., Duffy and Fehr, 2018) or asks each group to achieve a collective goal (e.g., Hortala-Vallve et al., 2013), it is appropriate to have a fixed group to interact over time. In this case, both individual-level clusters and group-level clusters can be inherited by the experimental design. If there are more than two ways of defining a cluster, and those ways are equally justifiable by the experimental design, then a researcher, given that he/she wants to report more robust statistical results, must choose a cluster within which observations are more related. One rule of thumb is to check the standard deviation of the observations within a potential cluster. For illustration, consider a public goods experiment with a fixed group of three subjects. Suppose that a researcher has collected data shown in Table 4.

	Group 1			Group 2		
ID	1	2	3	4	5	6
r01	0	10	5	2	5	3
r02	0	10	4	2	5	3
r03	0	10	4	3	5	1
r04	0	10	3	3	5	0
r05	0	10	3	3	6	0
r06	0	10	3	3	5	0
r07	1	10	3	3	5	0
r08	0	10	3	2	4	0
r09	0	10	3	1	4	0
r10	0	10	1	2	5	0
Std.	0.32	0.00	1.03	0.70	0.57	1.25
	Std.(Group)=4.25			Std.(Group)=1.95		

	Group 3			Group 4		
ID	7	8	9	10	11	12
r01	7	6	5	1	1	2
r02	3	4	4	5	3	3
r03	2	2	3	6	5	3
r04	0	1	4	3	5	4
r05	1	0	1	5	6	3
r06	0	1	0	7	5	6
r07	1	0	0	9	7	8
r08	0	0	0	10	10	9
r09	0	0	0	10	10	10
r10	0	0	0	10	10	10
Std.	2.22	2.07	2.06	3.17	3.08	3.19
	Std.(Group)=2.05			Std.(Group)=3.06		

Table 4: Strong dependence at the participant level (L) or the decision-group level (R)

If the individual choices vary little, as illustrated on the data from Groups 1 and 2 in Table 4, standard deviations of the participant-level observations (varying from 0.00 to 1.25) are smaller than those of the group-level observations (1.95 to 4.25). It implies that individual observations are more related to each other than group observations, so in this case, the standard error clustered at the individual level should be used. Meanwhile, if a group's choices vary less than individual choices, as illustrated on the data from Groups 3

and 4, the researchers may consider standard errors clustered at the independent-group level. I imagined situations where a group collectively reaches to complete free-riding or complete cooperation. Such a case may happen when group members' previous actions influence a subject's action more than the subject's own previous actions.[13]

The discussion above may be extrapolated to justify session-level clusters. If the session-level observations are more positively correlated than the individual's or the decision group's repeated choices, it could mean that the session-level clustered standard errors yield more robust statistical results. I am skeptical about this data-driven approach,[14] and I will discuss it after introducing cluster-robust inference in the following section.

4 Cluster-Robust Inference

In this section, I present a prototype parametric[15] model for cluster-robust inference of the mean treatment effect. I assume only one treatment (and one control) and that the experimenter controls session effects appropriately, so the model does not include them. An econometrician has $N = (S+S) \times I \times R$ observations in total, where S is the number of controlled and treated sessions, I is the number of per-session subjects, and R is the number of repetitions of the same game.[16]

For simplicity, set the dependent variable as the deviation from the mean control-group observations. Then, the treatment effect is captured by β in

$$y_i = \beta T_i + \varepsilon_i,$$

where $i = 1, \ldots, N$ is an index for observations, and $E[\varepsilon_i] = 0$. T_i has a value 1 if the observation is from the treated session and 0 otherwise. $\beta = 0$ implies that the treatment-group and the control-group means are the same. With a slight abuse of notation, T is a

[13] Some papers, e.g., Robbett (2014) and Gallo and Yan (2015), used the term "session" as a fixed independent group. In this case, it would be appropriate to cluster standard errors at the session (or independent-group) level for such cases.

[14] I should clarify that I do not mean to avoid any data-driven approach. Instead, I claim that the experimental design, or the intention of the researchers who design the experiment, should be prioritized over the purely statistical data features. If the experimental design well justifies two different clustering levels, then researchers could use a level that renders more (statistically) conservative reports. Thus I suggest observing some statistical features *within* the design-driven approach.

[15] Some researchers prefer non-parametric tests that take the session-level aggregate data as one independent data point. This approach may be free from the concern about the clustering issues as well as parametric assumptions, but the current paper does not address the comparative advantages of non-parametric methods.

[16] For expositional simplicity, I assume that the number of the subjects and the repetitions are the same for each session and that the number of controlled sessions is equal to the number of treated sessions, but these assumptions do not affect main messages.

set of treated observations such that for $i \in T$, $T_i = 1$. The OLS estimator is

$$\hat{\beta} = \frac{\sum_i T_i y_i}{\sum_i T_i^2} = \frac{\sum_{i \in T} y_i}{SIR},$$

and the variance of the estimator is

$$V[\hat{\beta}] = E[(\hat{\beta} - \beta)^2] = \frac{V[\sum_{i \in T} \varepsilon_i]}{S^2 I^2 R^2}$$

$V[\sum_{i \in T} \varepsilon_i] = \sum_{i \in T} \sum_{j \in T} Cov[\varepsilon_i, \varepsilon_j] = \sum_{i \in T} \sum_{j \in T} E[\varepsilon_i \varepsilon_j]$ is of our interest. If errors are uncorrelated, that is, $E[\varepsilon_i \varepsilon_j] = 0$ for $i \neq j$, it becomes $\sum_{i \in T} E[\varepsilon_i^2]$, and its sample analog, $\sum_{i \in T} (y - \hat{\beta} T_i)^2 = \sum_{i \in T} u_i^2$, yields the heteroskedasticity-robust standard error. We are concerning that this is not the case, at least within a cluster. Let C_i denote the cluster that i belongs to. If $E[\varepsilon_i \varepsilon_j] \neq 0$ for i and $j \in C_i$,

$$V_{clu}[\hat{\beta}] = \frac{\sum_{i \in T} \sum_{j \in T} E[\varepsilon_i \varepsilon_j] \mathbf{1}_{j \in C_i}}{S^2 I^2 R^2},$$

where $\mathbf{1}_A$ is an indicator whose value is 1 when condition A holds and 0 otherwise. Given that the number of clusters is sufficiently large,[17] we can use the variance estimate

$$\hat{V}_{clu}[\hat{\beta}] = \frac{\sum_{i \in T} \sum_{j \in T} u_i u_j \mathbf{1}_{j \in C_i}}{S^2 I^2 R^2}$$

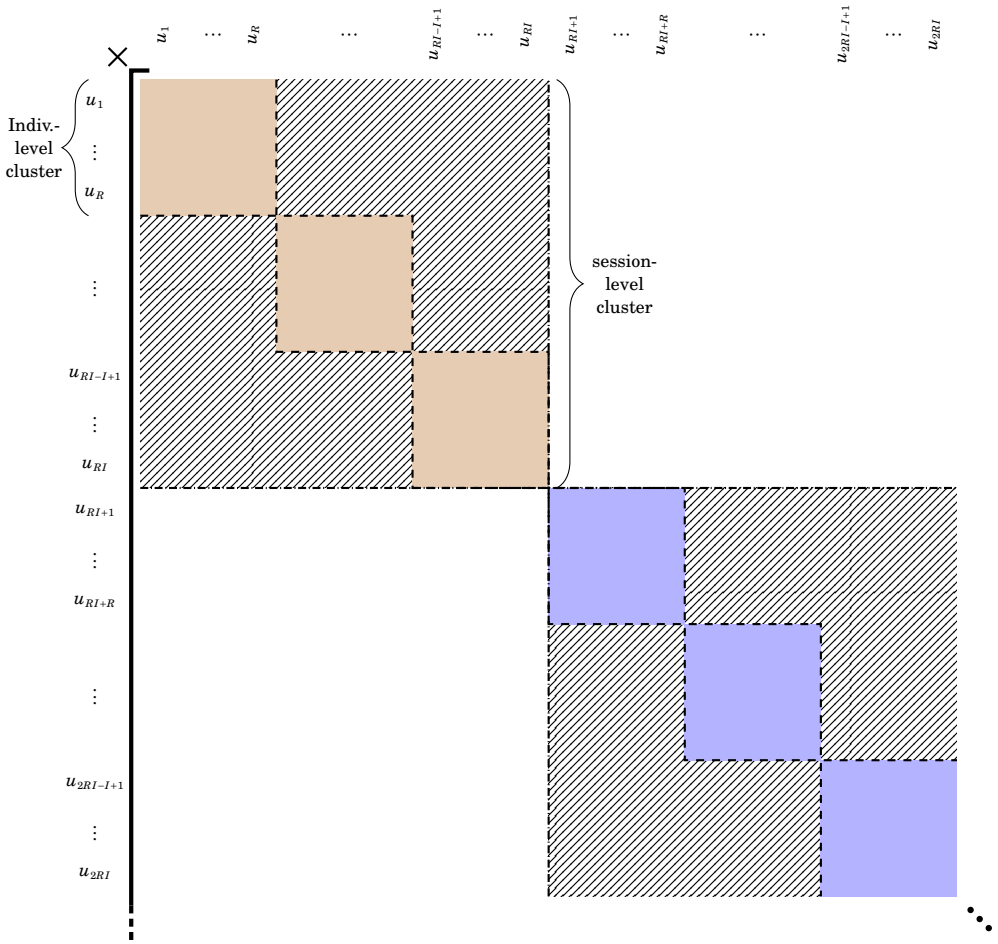
Two remarks are (1) if the cluster is the entire set, $\hat{V}_{clu}[\hat{\beta}]$ becomes zero because $\sum_{i \in T} u_i = 0$, and (2) if clusters are defined in a far-fetched manner so that $u_i u_j$ is negative for many pairs of i and j, the cluster-robust variance estimate could even be smaller than the heteroskedasticity-robust one.

Without loss of generality, lexicographically order the observations such that $i = s \times n \times r$, $s = 1, \ldots, 2S$, $n = 1, \ldots, I$, and $r = 1, \ldots, R$. Then $\sum_{i \in T} \sum_{j \in T} u_i u_j \mathbf{1}_{j \in C_i}$ is the summation of entities on the block diagonal matrices of $u_i u_j$, $i, j \in T$. Figure 2 illustrates the difference between standard errors clustered at the individual level and the session level. The main difference is that there are more off-diagonal (but still within a larger block diagonal) entities when clustering standard errors at the session level (see hatched areas in Figure 2). If standard errors clustered at the session level are larger than those at the individual level, it implies that the summation of those off-diagonal entities is positive. It happens when the signs of u_i and u_j are, in general, the same for $j \in C_i$.

[17] Another concern would regard the asymptotic refinement of the clustered standard errors when the number of clusters is small (Cameron et al., 2008). Kézdi (2004) shows simulation results that 50 clusters are often large enough for accurate inference. A typical laboratory experiment has much fewer sessions than 50.

Since the residual is the deviation from the conditional mean, the same signs imply error correlations.

Figure 2: Individual-level vs. session-level clusters



This figure illustrates a part of N-by-N matrix where entity at (i, j) is $u_i u_j$. The cluster-robust standard error of the treatment effect is the sum of the entities on block-diagonal sub-matrices. Clustering standard errors at the session level, compared to the individual level, involves more off-diagonal entities.

If the experimental design inherits the strong correlation between, for example, the first choice of individual i and the last choice of individual j in the same session, then the session-level cluster might be used. Perhaps someone's initial choice profoundly affects other's later choices so that those observations are related. Many questions can be followed. Is that relationship stronger than the relationship between a subject's own choices? Is that relationship stronger than the relationship between the last observations in one session and those in another session with the same treatment condition? It is undoubtedly possible that errors are weakly but positively correlated within a session, but considering a larger-size cluster comes at a price. Given the same number of obser-

vations, larger-size clusters have a more downward bias due to fewer clusters. Although statistical analysis software uses finite-cluster corrections,[18] it is unclear whether the standard error's downward bias will be appropriately corrected when a session is used as a clustering unit. While the experimenters may be concerned about the observational relationship within a session for any laboratory experiments, they should want to double-check whether the experimental design inherits the relationship from the beginning.

5 Simulations

For backing up the illustrations in Section 3, this section shows some Monte-Carlo simulation results.[19] For all simulation results, I consider $S = 4$ (four sessions per control and treatment each), $I = 18$ (18 subjects per session), and $R = 5$ (the last five repetitions of the game). Further, I assume that the group size is three (or six groups per session) and a random rematch (six groups are randomly shuffled every round). For this simulation, I have in mind a standard public goods game where a subject can choose a contribution level between 0 and 50 or a Tullock contest where a subject can invest up to 50 tokens to win the prize. Simulations are conducted in the following way.

1. Generate the treatment indicator, the session number, subject id, and the group number.[20]

2. For each iteration, $2 * S * I * R$ observations are generated in the following way.

   (a) In the first round of the experiment, each subject draws a choice from a discrete uniform distribution between 0 and 50.

   (b) From the second round and beyond, subjects tend to (i) stick to their previous choice and (ii) consistently respond to their previous group choices. Thus, the observation in the next round is a linear combination of three numbers: the number chosen by the subject in the previous round (with linear coefficient ρ_{ind}), the average number chosen by the group members in the previous round (with ρ_{ss}), and the randomly generated number from the same discrete uniform distribution (with $1 - \rho_{ind} - \rho_{ss}$).

3. Regress observations on the treatment dummy and a constant.

[18] For example, Stata uses $\frac{G}{G-1}\frac{N-1}{N-k}u_i$ instead of u_i, where G is the number of clusters, N is the number of observations, and k is the number of regressors.

[19] The code and the instructions are available at the Open Science Foundation repository.

[20] A fixed match is not considered in this simulation, but the simulation code can also serve the purpose. Check the instructions for the simulation.

4. Calculate the heteroskedasticity-robust standard error, the standard error clustered at the session level, and the standard error clustered at the individual level. Count if the p-value of the t-statistic ($\frac{\hat{\beta}}{SE}$) is less than 0.10.

5. Repeat Steps 2–4 for 1,000 times.

At least four points are worth mentioning. First, for this simulation, the population mean of the control is the same as the population mean of the treatment. Since the primary purpose of this exercise is to check the claim that the standard errors clustered at the session level may lead to a false-positive result (that is, reporting a statistically significant treatment effect when there is supposed to be no treatment effect), it is important to set no fundamental differences between the control and the treatment. Second, the sign of ρ_{ss} captures the direction of the responses to the previous observations from the matching group.[21] For example, a positive ρ_{ss} can be interpreted that the subject tries to imitate the previous average observation (such as conditional cooperation in the public goods game and learning the optimal investment level by observing other's investments), and a negative ρ_{ss} implies that the subjects deviate what the average players do (such as more free-riding after observing sufficient contributions from others and more investment than the average level to win the contest). Third, the magnitude of ρ_{ss} is naturally limited as the number of subjects increases. If the experiment adopts a perfect stranger match with sufficiently large subjects, whatever the subjects had learned from the previous game has nothing to do with the new game. When a session consists of 18 subjects and the size of a group is three, the probability of meeting at least one member of the previous group again is $2/17 \approx 0.1176$. With having this probability in mind, I vary β from -0.20 to 0.20. Whichever the sign of ρ_{ss}, the larger value implies the larger observational dependence across subjects within a session. Fourth, since the first-round data is generated from the discrete uniform distribution, and the second round and beyond depend on the initial realizations, a learning effect toward a particular decision point (for example, a Nash equilibrium) is not considered.

Table 5 shows the simulation results with different ρ_{ind} and ρ_{ss}. Three columns under "Mean" show the average value of heteroskedasticity-robust standard errors (SE_{het}), standard errors clustered at the session level (SE^{ss}_{clu}), and the standard errors clustered at the individual level (SE^{ind}_{clu}), respectively. The following three columns show the standard deviations of those standard errors, and the last three columns show the fraction of false-positive reports. Since there are no population differences between the control

[21] I assume that ρ_{ind} is always non-negative because it captures the subject's decision consistency. $\rho_{ind} \approx 0$ means that the subject merely ignores what he/she previously chose, and $\rho_{ind} < 0$ implies that the subject intentionally oscillates the decisions.

		Mean			St.Dev.			Pr(p-value<0.1)		
ρ_{ind}	ρ_{ss}	SE_{het}	SE^{ss}_{clu}	SE^{ind}_{clu}	SE_{het}	SE^{ss}_{clu}	SE^{ind}_{clu}	SE_{het}	SE^{ss}_{clu}	SE^{ind}_{clu}
0.8	-0.2	0.8636	1.2859	1.6645	0.0309	0.3757	0.0835	0.144	0.071	0.016
0.8	-0.1	0.8261	1.4347	1.6689	0.0305	0.4182	0.0768	0.176	0.069	0.032
0.8	0	0.8088	1.6489	1.7024	0.0304	0.4791	0.0714	0.215	0.072	0.046
0.8	0.1	0.8186	1.9520	1.7762	0.0313	0.5685	0.0702	0.243	0.072	0.063
0.8	0.2	0.8585	2.3552	1.8906	0.0339	0.6876	0.0757	0.286	0.069	0.089
0.5	-0.2	0.9201	1.0919	1.3252	0.0219	0.3242	0.0751	0.080	0.060	0.017
0.5	-0.1	0.8466	1.1289	1.2674	0.0211	0.3338	0.0703	0.105	0.060	0.027
0.5	0	0.7809	1.1893	1.2205	0.0209	0.3501	0.0658	0.149	0.068	0.041
0.5	0.1	0.7250	1.2858	1.1896	0.0214	0.3762	0.0619	0.189	0.074	0.059
0.5	0.2	0.6826	1.4335	1.1802	0.0226	0.4174	0.0590	0.221	0.070	0.085
0.2	-0.2	1.1069	1.0566	1.2323	0.0196	0.3109	0.0700	0.045	0.056	0.026
0.2	-0.1	1.0206	1.0608	1.1680	0.0182	0.3130	0.0662	0.053	0.063	0.037
0.2	0	0.9390	1.0717	1.1062	0.0173	0.3172	0.0627	0.068	0.061	0.046
0.2	0.1	0.8624	1.0924	1.0481	0.0168	0.3238	0.0597	0.092	0.061	0.059
0.2	0.2	0.7913	1.1281	0.9953	0.0167	0.3340	0.0569	0.131	0.063	0.072
0	0	1.0950	1.0567	1.0949	0.0182	0.3110	0.0616	0.045	0.058	0.045

Table 5: Monte-Carlo simulation results

and the treatment, such fraction should converge to 0 as the number of observations increases.

One can observe that the standard deviation of SE^{ss}_{clu} is distinctively larger than those of SE_{het} and SE^{ind}_{clu}. Figure 3 shows a histogram of one of the results ($\rho_{ind} = 0.8, \rho_{ss} = 0.1$) summarized in Table 5, which clearly illustrates that the standard errors at the session level (Clu-Ss) vary substantially more than heteroskedasticity-robust standard errors (Het) and the standard errors clustered at the individual level (Clu-Ind). It implies that the test statistics would substantially vary when the standard errors are clustered at the session level, although the data are obtained through an identical process. As a result, we observe more false-positive results when clustering the standard errors at the session level than clustering at the individual level, except for cases with $\rho_{ss} = 0.20$.

Another noticeable result from this simulation is that standard errors clustered at the session level are sometimes *less robust* than the heteroskedasticity-robust standard errors, especially with weak or negative dependence across observations. At the same time, it never happens with standard errors clustered at the individual level. It means that the attempt to find more robust statistical results could undesirably lead to the opposite outcomes when clustering standard errors at the session level.

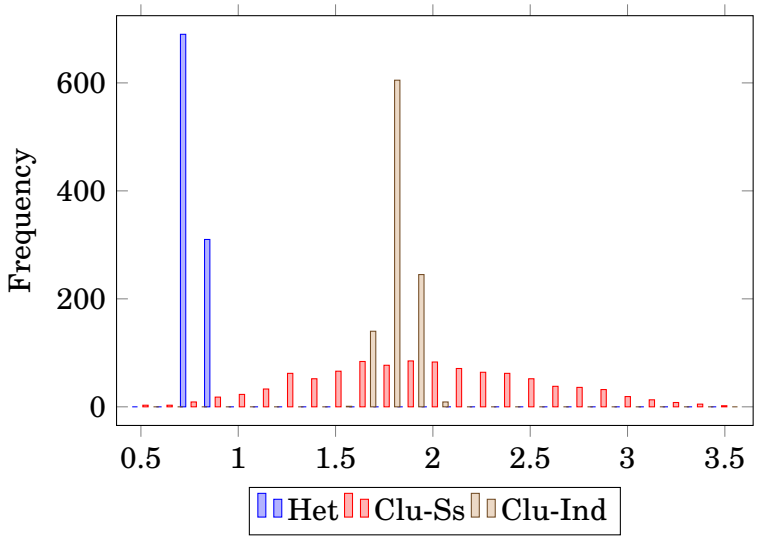


Figure 3: Simulated (iter=1000) standard errors clustered at the session level vary.

6 Discussions

6.1 Standard deviation as a rule of thumb

Suppose clustering observations can be done in two or more ways, equally justifiable by the experimental design. In that case, a researcher ought to choose a cluster within which the observations are more related to each other. I propose to check the within-cluster standard deviations of the observations. Recall that the residuals of the simple regression are the deviations from a conditional mean. A sufficiently smaller within-cluster standard deviation than the standard deviation of the entire sample may imply that the residuals flock together within sessions, and hence the errors are correlated within the cluster. Thus, when both session-level and individual-level clusters are equally justifiable by the experimental design, my rule of thumb is to compare within-cluster standard deviations. Consider I individual-level clusters, and S session-level clusters, where an individual-level cluster is a proper subset of a session-level cluster. Let std_I and std_S respectively denote the standard deviation of the individual-cluster observations and that of the session-cluster observations. If $std_I < std_S$ in general, then consider clustering the standard errors at the individual level.

If std_S is distinctively smaller than the standard deviation of the entire sample of the same treatment, then the session-level clustering might lead to larger standard errors. If it is the case, especially when the experimental design does not inherit the observational relationship within a session, a researcher may want to check whether the sessions are sufficiently randomized. A relevant situation is illustrated in Figure 4, which displays a scatterplot of observations from eight (four control and four treatment) sessions. Al-

most all residuals from sessions 1, 4, and 8 are positive, and almost all residuals from sessions 2, 3, and 6 are negative. Thus, the multiplications of those residuals within a session have positive values, and the standard error clustered at the session level will be larger than the heteroskedasticity-robust one. However, if thinking conversely, one may wonder whether the samples are balanced because otherwise, it is hard to explain the stark differences between identically-treated sessions. This distinctive variation across sessions may be due to the failure of session randomization or the session size being too small.

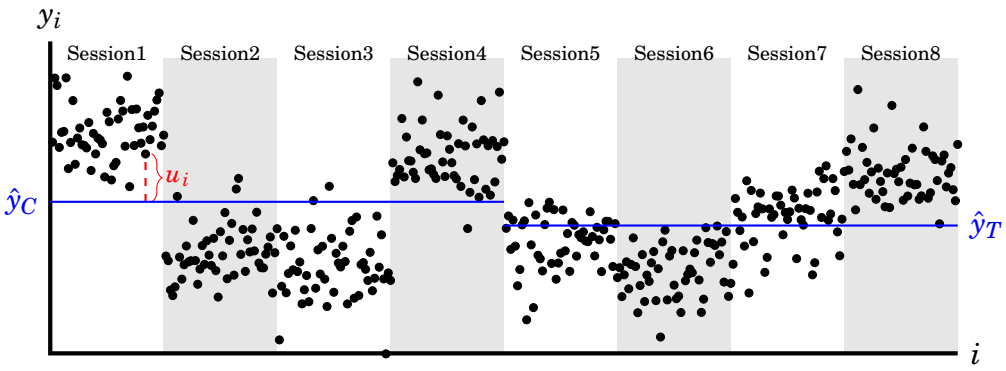


Figure 4: Small session-level standard deviations may question balanced sampling.

6.2 Further thoughts on the session-level clustering

I have claimed that the session-level cluster should be used either when the experimental design inherits the observational relationship within a session or when the session-level residuals flock together. The latter reasoning is inconclusive as it relies on the mechanical aspects of the data, not on the experimental design. If researchers consider clustering standard errors at the session level because it generates large standard errors, why not considering at the date-of-session or time-of-session level, why not at the experimenter level, and why not at the experiment level for meta-analysis[22] if those do the same or a better job? Furthermore, if we are willing to embrace let-the-data-tell-us approaches, why don't we consider bootstrapping or random forest algorithms?

An ad-hoc definition of a session also obscures the session-level clustering. Suppose 24 subjects show up at the lab, and the experimenter decides to split the subjects into two subgroups without informing them, but across the subgroups, the experiment proceeds identically. In this case, would a session consist of 24 subjects, or would two sessions

[22] Embrey et al. (2017) provide a meta-analysis of prior experimental research on the finitely repeated prisoner's dilemma and report the standard errors clustered at the study level. Detailed discussions and robustness checks on the clustering level for a meta-analysis are in the paper's Appendix A.4.

consist of 12 subjects each? This concern becomes more relevant to the fixed-group experiment. Suppose there are 24 subjects in one session, but 12 subjects show up in another session due to severe no-shows. If a fixed group of six subjects repeatedly play a game, one session is a cluster of four decision groups, and the other one is a cluster of two decision groups. If the former session's observations are seemingly less related because of more (potentially heterogeneous) groups, the session with fewer participants affects standard errors clustered at the session level more. Is having a different weight on each session justified?[23]

Another practical issue is the trade-off between the session size and the number of sessions. Given that the number of total participants is practically limited, considering session-level clusters pushes researchers toward having more sessions with fewer subjects per session. It is problematic in several aspects. First, many experiments adopt a random rematch design to minimize the strategic interactions between the games. If the subject size per session is small, then the indirect interactions are indispensable. If a subject plays ten games with a randomly paired partner in a session of 40 subjects, the probability that a subject does not meet any match again is 28.34%, but with 16 subjects per session, that probability plummets to 1.89%. Such a low probability implies that, with fewer subjects per session, the fundamental reason for adopting a random rematch is compromised: Although the subjects do not know whether the current match is new, they know that it is highly likely to have met before or would meet again. Second, fewer subjects per session can prevent us from having a balanced sample: Given that the subjects are drawn from the same population distribution, small-sized sessions feature more (un)observable variations in sessions.[24] Suppose each session consists of only four subjects each, and the female proportion dramatically varies from 0% to 100%. How can a researcher be sure whether the session effects are controlled, and if not, how does she distinguish the gender-ratio effect from others unless having more sessions with sufficiently large variations of the gender ratio? What is even worse, if the substantial variations across sessions are due to unobservable characteristics, not like observable gender ratios? A vicious cycle of demanding more sessions to control issues with small-size sessions may be established.

My argument here is simple. Suppose a researcher considers either 12 sessions with 6 subjects per session or 4 sessions with 18 subjects per session. If a researcher adopts a random rematch, intending that this design disconnects or minimizes the dynamic

[23] Moreover, Müller (2020) points out that when clusters are of different sizes, the p-values from typical statistical packages, such as Stata, are not reliable.

[24] Tversky and Kahneman (1974) point out that most people are insensitive that smaller samples are more likely to vary. If both a large and a small hospital recorded the days when more than 60% of the newborns were boys, which hospital is more likely to record more such days? Only 22% of the subjects correctly answered the small hospital.

relationships between decision rounds, then it is better to have 4 sessions not only because the standard errors clustered at the individual level would lead the false-positive treatment effect less but also because it is more aligned with the researcher's intention. However, if the researcher intends to facilitate the interactions across subjects over decision rounds, it would be better to have many (small-sized) sessions.

7 Conclusions

Session-specific idiosyncratic features can and should be integrated out when the researchers carefully randomize the sessions. If the purpose of clustering standard errors is to make more robust standard errors to minimize false-positive treatment effects, then one must consider clusters within which observations are more related, but across which observations vary. In a controlled laboratory experiment where participants repeatedly make choices in the same environment, individual-level clusters should be considered first, as the experimental design inherits the observational similarity within an individual. Takeaway messages are summarized below:

1. The experimenter's crucial responsibility is to ensure the participants in both the control and the treatment sessions are from the same population distribution and make each session environment as homogeneous as possible.

2. If the experiment asks participants to make repeated decisions in a similar environment, the experimental design inherits clusters at the participant (or independent decision-group) level. Thus, it is natural to cluster standard errors at the participant level.

3. The standard deviation of individual-level observations tends to be smaller than that of session-level observations. Thus, clustering standard errors at the participant level may yield more conservative statistical results.

4. If the experimental design equally justifies two ways of clustering observations, a researcher would choose a cluster within which observations are more related.

5. Although not justifiable by the experimental design, clustering standard errors at the session level may be considered if the session-level observations are related more strongly than those of participant- or group-level observations. It begs further questions of why a session should be a level for clustering, among several other potential levels, and whether the sessions have balanced samples.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge, "When Should You Adjust Standard Errors for Clustering?," Working Paper 24003, National Bureau of Economic Research November 2017.

_ , _ , _ , and _ , "Sampling-Based Versus Design-Based Uncertainty in Regression Analysis," *Econometrica*, 2020, *88* (1), 265–296.

Bracha, Anat, Uri Gneezy, and George Loewenstein, "Relative Pay and Labor Supply," *Journal of Labor Economics*, 2015, *33* (2), 297–315.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller, "Bootstrap-Based Improvements for Inference with Clustered Errors," *The Review of Economics and Statistics*, 2008, *90* (3), 414–427.

Carpenter, Jeffrey, "The labor supply of fixed-wage workers: Estimates from a real effort experiment," *European Economic Review*, 2016, *89*, 85–95.

Cipriani, Marco, Antonio Guarino, Giovanni Guazzarotti, Federico Tagliati, and Sven Fischer, "Informational Contagion in the Laboratory," *Review of Finance*, 06 2017, *22* (3), 877–904.

Corgnet, Brice, Mark Desantis, and David Porter, "What Makes a Good Trader? On the Role of Intuition and Reflection on Trader Performance," *The Journal of Finance*, 2018, *73* (3), 1113–1137.

de Chaisemartin, Clément and Jaime Ramirez-Cuellar, "At What Level Should One Cluster Standard Errors in Paired Experiments, and in Stratified Experiments with Small Strata?," Working Paper 27609, National Bureau of Economic Research July 2020.

Duffy, John and Dietmar Fehr, "Equilibrium selection in similar repeated games: experimental evidence on the role of precedents," *Experimental Economics*, 2018, *21*, 573–600.

Embrey, Matthew, Guillaume R. Fréchette, and Sevgi Yuksel, "Cooperation in the Finitely Repeated Prisoner's Dilemma," *The Quarterly Journal of Economics*, 08 2017, *133* (1), 509–551.

Engelmann, Dirk and Guillaume Hollard, "Reconsidering the Effect of Market Experience on the "Endowment Effect"," *Econometrica*, 2010, *78* (6), 2005–2019.

Ericson, Keith M., "Design Issues in Economics Lab Experiments: Randomization," 2018.

Fréchette, Guillaume R., "Session-effects in the laboratory," *Experimental Economics*, Sep 2012, *15* (3), 485–498.

Gallo, Edoardo and Chang Yan, "The effects of reputational and social knowledge on cooperation," *Proceedings of the National Academy of Sciences*, 2015, *112* (12), 3647–3652.

Hortala-Vallve, Rafael, Aniol Llorente-Saguer, and Rosemarie Nagel, "The role of information in different bargaining protocols," *Experimental Economics*, 2013, *16*, 88–113.

Kézdi, Gábor, "Robust Standard Error Estimation in Fixed-Effects Panel Models," *Hungarian Statistical Review*, 2004, *Special 9*, 96–116.

Moffatt, Peter G., *Experimetrics: Econometrics for Experimental Economics*, London New York, NY: Macmillan International Higher Education, 2016.

Müller, Ulrich K., "A More Robust t-Test," Working Paper 2020.

Robbett, Andrea, "Local Institutions and the Dynamics of Community Sorting," *American Economic Journal: Microeconomics*, August 2014, *6* (3), 136–156.

Tversky, Amos and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases," *Science*, 1974, *185* (4157), 1124–1131.