Clustering standard errors at the "session" level*

Duk Gyoo Kim†

May 27, 2020

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, which is crucial to deal with sampling issues, cannot justify clustering the standard errors at the session level. The experimental design 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, statistical inference can 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. A rule of thumb using standard deviations is introduced.

Keywords: Lab experiment, cluster-adjusted standard errors

JEL codes: C18, C90

1 Introduction

The purpose of this paper is 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 at the same time. For an experiment adopting a between-subject design, a subject participated in one session only,[1] so a set of observations from an individual is a proper subset of the

*I thank Guillaume Fréchette, Franziska Heinicke, Sang-Hyun Kim, Wooyoung Lim, Yoshiyasu Rai, Euncheol Shin, and Wladislaw Mill, and participants at Mannheim/ZEW Experimental seminar for their helpful comments, and Elisa Casarin for her research assistance.

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

[1]On the contrary, a within-subject design assigns a participant to two or more treatments. In this case, clustering standard errors at the session level is less persuasive as the primary purpose of the design is to examine individual changes.

whole observations from a session, which is a proper subset of the whole observations from the same treatment. For this reason, adding individual or session fixed effects to the regression does not help to 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), 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 the reason 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 explicitly mentioned that cluster-robust standard errors are used. Standard errors of 40 papers are clustered at the participant level, and those of 34 papers are clustered at the session level.[3]

I often find arguments that standard errors should be clustered at the session level when the session-specific effects are concerned. 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] On top of this, 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 session effects in the laboratory (Fréchette, 2012). This reasoning—using session-level cluster adjustment for session effects—is not on solid ground. A concern for session effects is the reason for randomizing the sessions as much as possible so that the session-specific idiosyncratic features can be integrated out, not for clustering standard errors at the session level. I am worried that many researchers seem to insist session-level clustered standard errors as a remedy of session effects, without further justifying why a session should be the cluster level. As I will elaborate later, the clustering level should be determined by the experimental design,[6] and when the design inherits observational relationship within a session, standard er-

[2] Moffatt (2016) explains that different (subject-level as the lowest and session-level as the highest) clustering can be considered. 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 it in Section 3.

[4] In a typical laboratory experiment, one session is conducted at one time, so session-level randomization practically implies the random ordering of the sessions.

[5] More details can be found in his blog post (Ericson, 2018, link).

[6] Abadie et al. (2020) claim to consider design-based uncertainty, as opposed to sampling-based one, for statistical inference. My claim is in line with theirs.

rors should be clustered at the session level. Figure 1 summarizes my arguments.

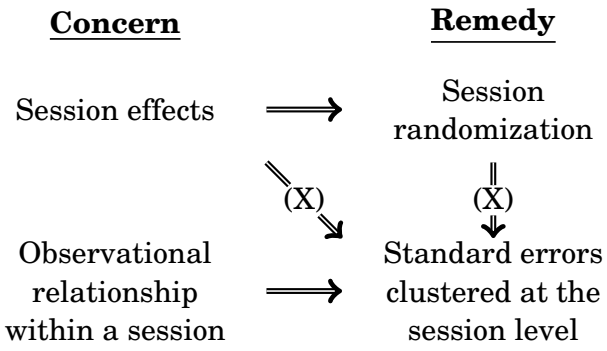


Figure 1: Clustering at the session level is not a remedy for session effects.

To minimize destructive discussions, I emphasize two things that I am *not* claiming. First, I am not claiming that there are no session-specific effects. It is the experimenter's crucial responsibility to maintain the environment of every session 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 and the treatment-group participants are from the same population by randomizing the order of the sessions. 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 to conduct more sessions if they believe potential session effects should be controlled. For the same reason, a sequential modification of the experimental design—for earlier sessions conducting X and Y and for (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 session effects, it is a reason for being careful about sampling subjects from the same population pool by randomizing the sessions 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.

A session-level cluster can undoubtedly address the observational dependence within the session, and 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.[7] Although Fréchette (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),[8] 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, not a negation of it. Again, the current paper focuses less on the studies where the experimental design inherits observational relationship within a session, which I strongly believe the session-level clustering is appropriate, but it focuses on the discussions about the proper cluster level when individuals in the lab make repetitive decisions of the same game.

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 may be considered carefully. Section 4 presents a simple econometric model to address issues related to the session-level cluster. Section 5 summarizes the lessons from this project, and Section 6 concludes.

2 Why do we adjust standard errors at a cluster level?

Clustered standard errors should be considered when observations within a cluster are related to each other. In words, if the observations within a cluster are similar, then the errors within a cluster will be more correlated than those of the whole observations. Thus, without "penalizing" the observational similarity, we will have downward-biased standard errors, which may lead to false-positive treatment effects more often.

To elaborate on what I mean by "penalizing" similarity, consider the following situation. There are two sets of five observations: one set from a control group experiment, and the other set from the treatment group experiment. Assume further that other than the treatment condition, everything else is homogeneous and appropriately controlled. A researcher wants to examine if the mean control-group observation is statistically dif-

[7] For example, Engelmann and Hollard (2010) have participants made only a small number of decisions and focus more on the interaction within a session. The main interest of Cipriani et al. (2017) is on the session-level information contagion, so the interactions within a session is 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 possible gains of other traders within the 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.

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

ferent from the mean treatment-group observation.

	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 below the original entities.

	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 t-test rejects the null hypothesis (p-value=0.0487). The (un-clustered) standard error of the mean difference is 0.040. 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 conservative statistical results.

A naturally followed question is what would be the proper cluster level. Unlike other fields of studies where the clustering units can be multi-dimensional, potential clusters in a between-subject experiment—individual or session—are uni-dimensional. That is, 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 When is the session-level clustering more robust?

If standard errors cluster-adjusted at the session level are larger than those at the individual level, then 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 or similar environment repeatedly.

To support my claims, I use hypothetical data instead of actual data from the previously published studies. 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,[9] 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 repetitive decision rounds choosing an integer between 5 and 45, 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 that a researcher has collected data from four (two control and two treatment) sessions, as shown in Table 2.

[9]For a decision-theoretic lab experiment, that is, 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 that involve strategic decisions.

	Control-Session-01						Treatment-Session-01						Control-Session-02						Treatment-Session-02					
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

Each column of Table 2 is a vector of an individual's decisions over ten rounds. A 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 less than 2.7, which implies that individuals made similar choices over rounds. The session-level standard deviations, around 11, are as large as the standard deviation of the whole observations. If we cluster the standard error of the mean difference at the participant level, the difference is no longer statistically significant (see Table 3). However, clustering standard errors at the session level does not help us to handle the false-positive treatment effect.

	(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 deci-

sions,[10] the observational similarity at the participant level is *inherited from* the experimental design when the experiment asks a participant to make repetitive decisions in the same environment. 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 dynamic changes of the actions are not of their primary interest, researchers often focus on analyzing the observations from the later half decision rounds. Those observations are likely to be "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 repetitive 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 the independent groups to achieve a collective goal (e.g., Hortala-Vallve et al., 2013), then 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 at the decision-group level (R)

If the choices of an individual vary little, as illustrated on the data from Groups 1 and 2 in Table 4, then standard deviations of the participant-level observations (varying

[10] In this case, the incentive structure of the experiment should be criticized.

from 0.00 to 1.25) are smaller than those of the group-level observations (1.95 and 4.25). In this case, the standard error clustered at the individual level should be used. Meanwhile, if the choices of a group vary less than individual choices do, as illustrated on the data from Groups 3 and 4, then the researchers may consider standard errors clustered at an 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 have influenced a subject's action much more than the subject's own previous actions.[11]

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

4 Cluster-Robust Inference

In this section, I present a prototype parametric[12] model for cluster-robust inference of the mean treatment effect. I assume that there is only one treatment (and one control) and that the experimenter controls session effects well, 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 subjects per session, and R is the number of repetitions of the same game.[13]

For simplicity, set the dependent variable as the deviation from the mean control-group observations. Then, the treatment effect that we want to examine 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 mean

[11] Some papers, e.g., Robbett (2014) and Gallo and Yan (2015), used the term "session" as a fixed independent group. In this case, clustering at the session level is appropriate if the experiment focuses on the interactions among subjects within the group.

[12] Some researchers prefer non-parametric tests such as Mann-Whitney test and Wilcoxon signed-rank test, with taking 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 comparative advantages of non-parametric methods.

[13] 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.

treatment-group observation is the same with the control-group mean. With a slight abuse of notation, T is a 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 the middle part of the "sandwich matrix." If errors are uncorrelated, that is, $E[\varepsilon_i \varepsilon_j] = 0$ for $i \neq j$, then it simply becomes $\sum_{i \in T} E[\varepsilon_i^2]$. We are concerning that this is not the case, at least within a cluster. Let C_i denote the cluster (a partition of the entire observations) that i belongs to. If $E[\varepsilon_i \varepsilon_j] \neq 0$ for i and $j \in C_i$, then

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

where $\mathbf{1}_A$ is an indicator function whose value is 1 when condition A holds, and 0 otherwise. Given that the number of clusters is sufficiently large,[14] we can use the variance estimate

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

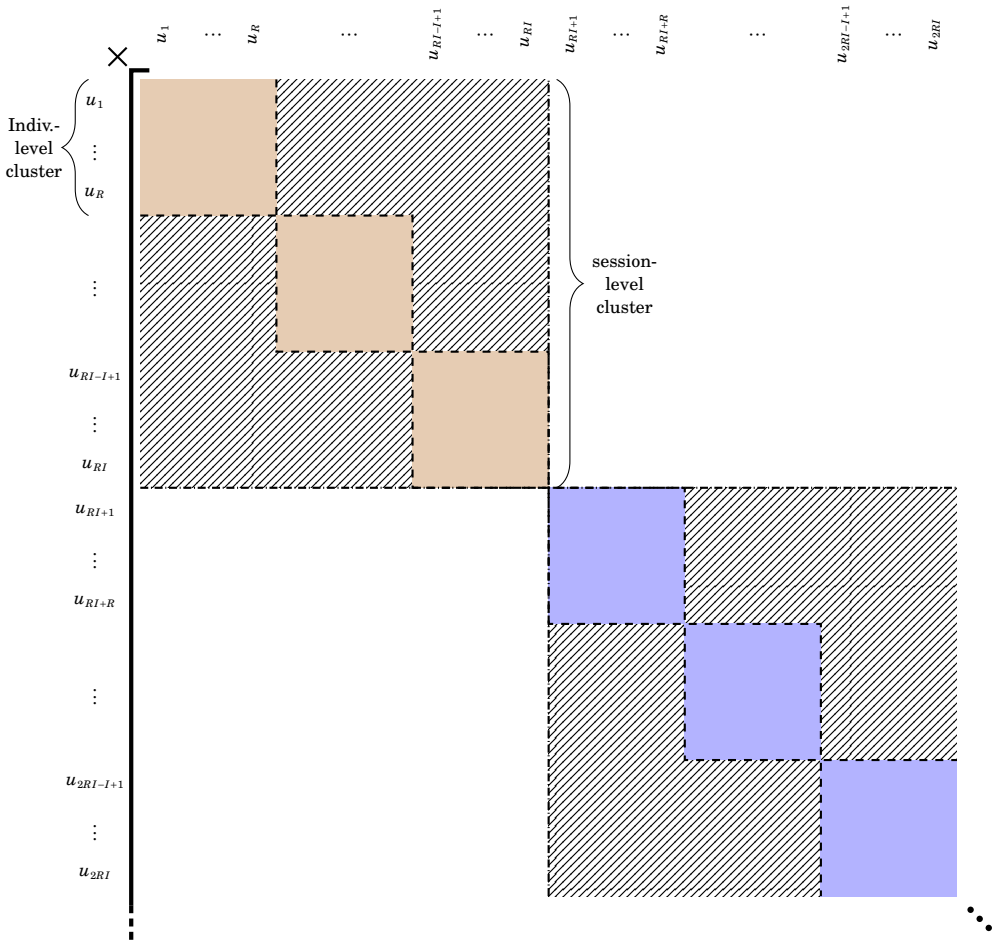
where $u_i = y_i - \hat{\beta} T_i$. It is worth noting that (1) if the cluster is the entire set, then $\hat{V}_{clu}[\hat{\beta}]$ becomes zero because $\sum_{i \in T} u_i = 0$, and (2) if clusters are defined in a far-fetched manner so that there are many pairs of i and j such that $sgn(u_i) \neq sgn(u_j)$, then the sign of $u_i u_j$ is negative, and the estimate could even be smaller than the heteroskedasticity-robust variance of the estimator, $\hat{V}[\hat{\beta}] = \sum_{i \in T} u_i^2 / (SIR)^2$.

Without loss of generality, lexicographically order the observations in a way 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, then 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 $j \in C_i$.

[14] Another question would regard the asymptotic refinement of the cluster-adjusted standard errors when the number of clusters is small (Cameron et al., 2008). I discuss it in Section 5 regarding the session-level clustering.

With recalling that the residual is the deviation from the conditional mean, the same signs imply that those are correlated.

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, say, the first choice of individual i and the twentieth choice of individual j in the same session, then the session-level cluster should 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? It is, of course, 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 observations, larger-size clusters have a

more downward bias due to the smaller number of clusters. Although statistical analysis software uses finite-cluster corrections,[15] it is unclear whether the downward bias of the standard error will be appropriately corrected when a session level is used as a clustering unit. While the experimenters may be concerned about the observational relationship within a session for any type of laboratory experiments, they may want to double-check whether the experimental design inherits the relationship from the beginning.

5 Discussions

5.1 Standard deviation as a rule of thumb

If the primary purpose of clustering standard errors is to provide a more robust treatment effect, and clustering observations can be done in two or more ways, which are equally justifiable by the experimental design, then a researcher chooses 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 compared to the standard deviation of the whole observations may imply that the residuals flock together, and hence they are correlated within the cluster.

Consider the following example. Suppose u_i, $i = 1, 2, \ldots, 100$, has either a value of $+1$ or -1 such that $\sum_{i=1}^{100} u_i = 0$. Then the (population) standard deviation is 1. A researcher considers two clusters of 50 values each. If $\sum_{i=1}^{50} u_i = \sum_{j=51}^{100} u_j = 0$, the standard deviation of each cluster is 1 and the sum of block-diagonal entities is zero. If $\sum_{i=1}^{50} u_i = 50$, which means that all $+1$'s are in the first cluster and -1's are in the second cluster, then the standard deviation of each cluster is 0 and the sum of block-diagonal entities is 2,500 each. Note that the cluster-robust standard errors are proportional to the sum of the block diagonal matrices. Having those two extreme cases in mind, one can easily check that the negative relationship between the within-cluster standard deviation and the magnitude of the cluster-robust standard errors (see Figure 3). In this example, if $|\sum_{i=1}^{50} u_i| \leq 6$, then the cure is worse than the disease, that is, the cluster-robust standard error is smaller than the heteroskedasticity-robust standard error.

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

[15] 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.

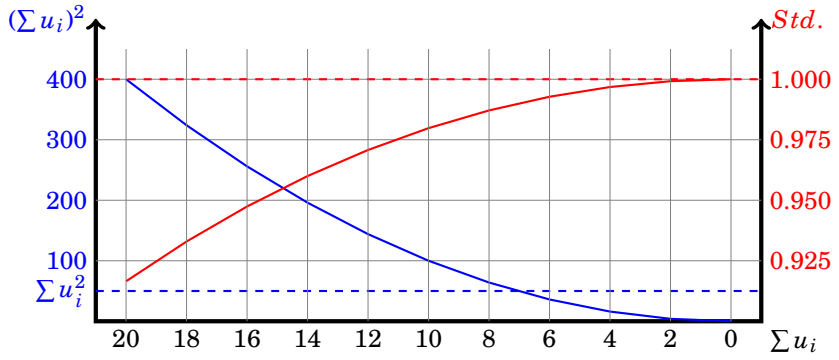


Figure 3: Within-cluster standard deviations (red) and cluster-robust variance (blue)

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 denote the standard deviation of the individual-cluster observations and the standard deviation of the session-cluster observations, respectively. If $std_I < std_S$ in general, then consider clustering standard errors at the individual level.

If std_S is distinctively smaller than the standard deviation of the whole observations of the same treatment, then the session-level clustering might lead to larger standard errors. However, if it is the case, a researcher may want to check whether the sessions are sufficiently randomized, especially when the experimental design does not inherit the observational relationship within a session. A relevant situation is illustrated in Figure 4, which displays a scatterplot of observations from eight (four control and four treatment) sessions. Almost 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. If thinking conversely, however, one may wonder whether the samples are balanced because otherwise, it is hard to explain the differences between sessions of the same treatment. This may be due to the failure of session randomization or the session size being too small.

5.2 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 second reasoning is not conclusive as it relies on the mechanical aspects of the data, not on the experimental design. If a researcher considers clustering standard errors at the session level because it generates larger standard errors, why not considering at the date-of-session or time-of-session level, why not

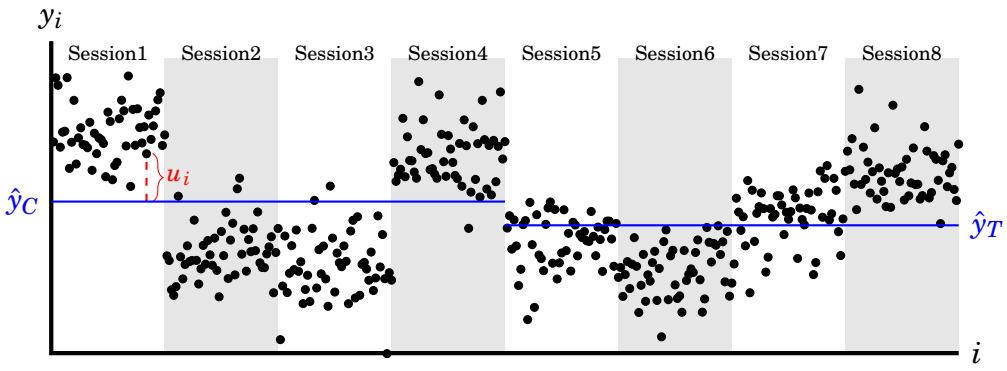


Figure 4: Small session-level standard deviations may question balanced sampling.

at the experimenter level, and why not at the experiment level for meta analysis[16] if those do the same job?

An ad-hoc definition of a session also makes the session-level clustering unclear. Suppose that there are 24 subjects show up at the lab and that the experimenter decides to split the subjects into two subgroups of twelve without explicitly informing them, but across the subgroups, the experiment proceeds in an identical manner. In this case, would a session consist of 24 subjects, or would two sessions consist of 12 subjects each? This concern becomes more relevant to the fixed-group experiment. Suppose there are 24 subjects in one session, but due to severe no-shows 12 subjects show up in another session. If a fixed group of six subjects played a game repeatedly, one session is a cluster of four decision groups, and the other one is a cluster of two decision groups. If the observations from the former session are seemingly less related because there are more (potentially heterogeneous) groups, then clustering standard errors at the session level will be more affected by the session with a smaller number of participants. Is having a different weight on each session justified?

Another practical issue is on the trade-off between the session size and the number of sessions. Given that the number of total participants is limited because of either the financial reason or the capacity of the subject pool, considering session-level clusters encourages researchers to have more sessions with a smaller number of 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 (Andreoni and Croson, 2008). 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, and there are 40 subjects per session, then the probability that a subject does not meet any

[16] 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 can be found in their Appendix A.4.

match again is 28.34%, but with 16 subjects per session, that probability plummets to 1.89%. Such a low probability implies that, with a smaller session size, 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 that they have met before or would meet again. Second, as I illustrated in the previous subsection, the small sample size per session can prevent us from having a balanced sample.[17] Suppose that each session consists of 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 if the substantial variations across sessions are due to unobservable characteristics, unlike the gender ratio? A vicious cycle of demanding more sessions to control issues with a small-size session may be established.

The following thought experiment can further clarify what I claim regarding the session size. Suppose a researcher has an unlimited physical, financial, and technical capacity to run a large-scale controlled experiment with 1,000 participants, and she considers two options. One option is to bring all participants at the same time, randomly divide them into two–one control and one treatment–groups, and identical twin experimenters conduct the experiment in two similarly-designed nearby places. The other one is to bring ten subjects at a different time and date, randomly order sessions, with hoping that the samples are balanced in aggregate and that the idiosyncratic session effects are integrated out, and the identical twin experimenters (with different levels of fatigue) conduct 50 sessions[18] of the experiment each. Insisting to cluster standard errors at a session level pushes researchers to choose the second option, which I find less appealing.

6 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 a cluster within which observations are more related, but across which observations are varying. In a controlled laboratory experiment where participants make

[17] Tversky and Kahneman (1974) point that a majority of people are insensitive to the fact that observational variation is more likely in smaller samples: When both a large hospital and a small hospital recorded the days on which more than 60% of the babies born were boys for a period of one year, only 22% of the subjects correctly answered that the small hospital is more likely to record more such days.

[18] Kézdi (2004) shows simulation results that 50 clusters is often large enough for accurate inference. Note that a typical laboratory experiment has a smaller number of sessions.

choices in the same environment repeatedly, individual-level cluster adjustment can be justified as the experimental design inherits the observational similarity within an individual. Takeaway messages are summarized below:

1. It is the experimenter's responsibility to ensure the participants in both the control and the treatment sessions are from the same population distribution.

2. If the experiment asks participants to make repetitive decisions in a similar environment, the experimental design inherits clusters at the participant (or independent decision group) level. Thus, it is natural to consider clustering 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 will yield more conservative statistical results.

4. If there are more than two ways of clustering observations and those are equally justifiable by the experimental design, then 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 standard deviation of the whole session-level observations is smaller than that of participant- or group-level observations. It begs a further question of why a session should be a level for clustering, among several other potential levels, and of 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.

Andreoni, James and Rachel Croson, "Partners versus strangers: Random rematching in public goods experiments," *Handbook of experimental economics results*, 2008, *1*, 776–783.

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.

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.

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.