When should we cluster experimental standard errors?[*]

Thomas Robinson[†]

April 21, 2020

Abstract

When researchers suspect that error terms are correlated by group in observational research the standard correction is to cluster the standard errors. But what about in experimental contexts where treatment is randomised? Despite their ubiquity in analyses with group-constant variables, the rationale for using clustered standard errors in experimental contexts remains underdeveloped. In this paper I present an intuitive and applied explanation of when clustering is appropriate, building on recent contributions in the statistics and econometrics literatures. I demonstrate why randomisation does not lead to identical variance estimates across estimation strategies, and conduct a review of experimental studies published between 2017 and 2019 to show that these differences can be considerable. Finally, I provide practical guidance for when and why to cluster standard errors for common experimental designs.

[*] This working paper forms part of my doctoral thesis "Three Essays on Measuring Political Behaviour".

[†] Department of Politics and International Relations, University of Oxford. Email: thomas.robinson@politics.ox.ac.uk

Standard errors help indicate the probable bounds of a point-estimate under the assumption of repeat sampling. We use these bounds to limit the likelihood of committing a Type I error (wrongly rejecting a true null hypothesis), and thus to assess our confidence over a given statistical finding. Standard errors are integral to our practice as experimentalists. Without them point-estimates of the average treatment effect (ATE) from finite samples are meaningless. We would not know the precision of the ATE and, crucially, whether it is distinguishable from a null effect.

Despite their foundational importance to statistical analysis, guidance for applied researchers on *which* estimator of uncertainty to use is still developing. Conventional estimates of variance are known to have serious limitations – for instance when observations in a dataset are clustered (Moulton, 1986; Liang and Zeger, 1986) or when residuals are heteroskedastic (Huber, 1967; Eicker, 1967; White, 1980). But we know less about when to justify the inclusion of modified variance estimators in our analysis, particularly outside of observational research contexts (Blattman, 2015; Abadie et al., 2020). Too often researchers seem to default to whichever variance estimator provides the most conservative (widest) uncertainty bound. The problem with this approach is that researchers often neglect that variance is a quantity to be estimated. Choosing the wrong estimator leads to incorrect inferences.

Neglecting the estimation of uncertainty is particularly acute in experimental research. Advances in the design of both in-the-field and online experimental apparatus have enabled scholars to estimate causal point-estimates across a wide variety of topics. The experimental method allows researchers to assume away the problem of selection bias when treatment assignment is randomised. But what about bias in the standard error? Does randomisation allow us to assume away potential biases of the corresponding variance estimates too?

There are numerous contexts in which it appears that experimental researchers should use some non-standard variance estimation strategy. In particular, observations within experimental research often involve multiple observations within groups. When individuals take part in repeat-observation tasks, randomised control trials (RCT) measure multiple observations within a "treated" geographical area, or researchers leverage some natural randomisation, there is likely to be a group-structure to statistical models' error terms. Randomisation of treatment in these protocols ensures regression coefficients are unbiased. But despite the temptation to think that this also solves the problem of clustered data, randomisation of treatment does not guarantee that errors terms will be uncorrelated within groups (Abadie et al., 2017).

Developed in the 1980s (Moulton, 1986; Liang and Zeger, 1986), cluster-robust estimates of variance continue to be refined and compared in the statistical and applied literatures (Esarey and Menger, 2019; Jackson, 2019; D'Adamo, 2019). It is now commonly accepted that failing to account for clustering can lead to downwardly biased standard errors increasing the likelihood of a Type I error (Wooldridge, 2003).

The incorrect application of clustering estimators, however, can itself result in biased standard errors. While the conventional expectation is that cluster-robust estimation inflates variance compared to conventional estimation (Wooldridge, 2003; Cameron and Miller, 2015), studies show that clustering can downwardly bias variance estimates too (Esarey and Menger, 2019; Jackson, 2019). Even when clustered standard errors are larger than their conventional counterparts, they may nevertheless substantially underestimate the true variance (Imbens and Kolesar, N.d.).

Fewer works, however, have examined variance estimation issues from the perspective of randomised experiments specifically (Green and Vavreck, 2008; Schochet, 2013; Cameron and Miller, 2015; Abadie et al., 2017). In particular, there is little specific guidance for experimental researchers on whether to use modified variance estimators when the data generating process includes some group-level, clustered component. Those that have considered this issue have offered disparate and often conflicting advice. While Cameron and Miller (2015) argue that clustering is inconsequential in experimental contexts due to randomisation, Abadie et al. (2017) argue that the decision to cluster is a "design-based" decision dependent on both how observations are sampled and how treatment is assigned. When should experimental researchers, therefore, cluster standard errors?

In this paper, I distil the latest research from the statistics and econometrics literatures on variance estimation (Cameron and Miller, 2015; Abadie et al., 2014, 2017, 2020) to provide an intuitive and practical guide to when and why experimental research designs should use clustered standard errors. In particular I make three contributions.

First, I provide a simplified argument about when to use cluster-robust variance estimation in experimental contexts. Building on the design-based approach to clustering developed by Abadie et al. (2017), I show how both sampling frames and treatment assignment matter for correct variance estimation. In short, researchers should cluster experimental errors when either sampling is done at the group-level (and one wants to infer to groups beyond the data) or treatment randomisation is intentionally correlated within groups.

Second, I demonstrate the implications of this argument using Monte Carlo simulations and a comprehensive literature review of experimental studies published in the *Journal of Experimental Political Science*. I show that even in completely randomised experimental designs there are likely to be considerable differences in the magnitude of conventional and cluster-robust standard errors. Correcting variance estimations in three published articles, moreover, I demonstrate that in practice these errors can be more than twice their original reported size.

Third, I emphasise how making valid statistical inferences in experimental research has to focus on both $\hat{\beta}$ and $var(\hat{\beta})$. One tempting approach to variance estimation is simply to report whatever variance estimator produces the largest standard error. We do so to hedge our bets against committing a Type I error. But this "largest error" approach makes the implicit judgement that a Type II error (failing to reject a false null hypothesis) is the lesser statistical evil compared to its Type I counterpart. This paper resists this claim on two fronts. From a statistical perspective, without considering the underlying data generating process, cluster-robust estimates may not be a conservative estimate of the true variance.[1] From a hypothesis testing perspective, correctly estimating variance (irrespective of its comparative size compared to other estimators) is an intrinsic part of making unbiased, scientific claims. A Type II error should concern researchers just as much as a Type I error. Over-estimating variance can lead to "file-drawer" problems – where true null results go unreported, biasing our research findings – if incorrect variance estimation leads to overly cautious estimates of standard errors.

The remainder of this paper proceeds as follows. Section 1 provides an account of why uncertainty matters, and the different forms it can take, in experimental research. Section 2 motivates the design-based logic of Abadie et al. (2017) to provide a simple, principled and easily-implementable variance estimation strategy for experimentalists. Sections 3 and 4 then demonstrate the importance of correctly estimating variance using Monte Carlo simulations and replication data respectively. Finally, Section 5 provides concrete guidance for two common types of experiment – randomised controlled-trials and conjoint experiments.

1 What is uncertainty in an experimental context?

Suppose an experimentalist wants to estimate a simple parameter like the average treatment effect (ATE; τ_{ATE}). To do so, they draw a series of observations – from a convenience sample like a student subject pool, or from a representative sample of a national population – and randomly assign individual units to treatment and control. Taking the mean difference in outcome between these two groups of observations, the researcher recovers an *estimate* of the ATE ($\hat{\tau}_{\text{ATE}}$).

Estimation implies uncertainty. It is an approximation of the true value. What the "true value" corresponds to, and therefore the extent of uncertainty over an estimate, will depend

[1] Note for instance that Cameron and Miller (2015) argue that in contexts where there is potential clustering one should compare conventional and cluster-robust variance estimators and opt for the cluster-robust estimate if there is an 'appreciable difference' (p.17). But this is only the correct strategy if we know that errors are in fact clustered at some group level and, moreover, that the separate assumptions for valid clustered variance estimation hold for the given data.

on how the researcher wishes to generalise their findings. From a very narrow perspective, the researcher may claim their is no uncertainty over their estimate. Given the specific set units in the sample (S), their assignment to treatment and control (Z), and the experimental context (E), we can calculate the average effect on the outcome exactly. In other words, conditional on the sample, randomisation, and experimental conditions, the recovered ATE estimate is the true value conditional on those same features:

$$E[Y(1) - Y(0)|S, Z, E] = \hat{\tau}_{\text{ATE}} = \tau_{\text{ATE}|S,Z,E} \tag{1}$$

Researchers, however, typically do not want to simply estimate this conditional effect and instead wish to generalise their result. In other words, researchers want to drop at least one of S, Z, and E from their claims; doing so is equivalent to generalising over a wider population, different treatment assignment schedules, and or contexts.[2]

$$E[Y(1) - Y(0)|S, Z, E] = \hat{\tau}_{\text{ATE}} \approx \tau_{\text{ATE}} \tag{2}$$

Generalisation involves uncertainty because we try to infer the parameter value for data which the researcher does not observe. Note in Equation 2 that the same estimate of the ATE is only an approximation of the unconditional ATE. Typically we are concerned with uncertainty that arises due to finite sampling issues. Out of the universe of cases (the population), researchers are only about to collect data on a small proportion of these (the sample). The uncertainty of a parameter estimate corresponds to the fact that we wish to make a generalisable claim about the population from the data in our sample. In other words, we want to make a claim that includes units $s' \notin S$.[3] Since in practice the researcher

[2] E captures features of the experimental environment – the mode or lab conditions, general experimental protocols, and other contextual factors – that condition how the experiment was conducted. Generalising over E captures external validity concerns – how well does the parameter estimate approximate the true effect in other experimental or non-experimental contexts. For instance, would a causal effect of an information prime delivered on a screen in a lab hold in a "real-world" context like a get-out-the-vote field experiment? For a more in-depth discussion of replicating effects across experimental contexts see, for example, Duch et al. (2020).

[3] For instance, consider the set of numbers $3, 2, 4, 9, 1$. The true mean of the set is 3.8. If we only ever

typically only samples from the population once, uncertainty over the estimate quantifies the probable bounds of the parameter given those units not in the sample.

This "sampling uncertainty" is the basis of variance and standard error calculation in most statistical research. However, it is just one particular form of uncertainty as a consequence of finite data. Various other factors may not be fully observed as a result of the one-shot nature of data collection, including how the estimate of the average treatment effect might differ if the treatment assignment schedule (Z) were different.

From a vast number of equally-probable assignment schedules we only observe one particular draw (Abadie et al., 2014). The notion of "causal" uncertainty, where there are many potential assignments, is the foundational basis of Neyman's randomisation inference (Neyman, 1990) yet receives less attention in current experimental research. Given a distribution of potential treatment assignments and resultant outcomes, how outlying and therefore significant is the *observed* difference between treatment and control? The substantive concern is, holding constant the subjects within our sample but re-randomising them into new treatment and control groups, how much would we expect the observed effect to differ? This question is inherently counterfactual. Unlike in the case of sampling uncertainty, where it is plausible to draw new observations from the population, we cannot draw new, unconfounded assignment schedules for the same individuals because they would already have been exposed to one treatment arm or another.

The counterfactual nature of this uncertainty does not negate the importance of it to researchers interested in estimating the true causal effect of a stimulus. To see why, consider the potential outcomes in Table 1. Note that all individuals perform uniformly except individual ($i = 3$) who always reports the same outcome regardless of the assigned treatment condition.

The true average treatment effect is $\frac{1}{M}\sum_{i=1}^{M} Y_i(1) - Y_i(0) = 4.2$. Suppose that the randomly drawn assignment vector $Z = \{1, 1, 1, 0, 0, 0\}$. Our best, unbiased estimate of the

sample two values our estimate will range from 1.5 ($\frac{1+2}{2}$) to 6.5 ($\frac{4+9}{2}$).

Table 1: Potential outcomes for six individuals under treatment and control conditions

i	$Y_i(1)$	$Y_i(0)$
1	10	5
2	10	5
3	**10**	**10**
4	10	5
5	10	5
6	10	5

ATE is the difference in means between those actually treated and those actually in control, $\overline{Y_{i1}}(1) - \overline{Y_{i0}}(0) = 10 - 5 = 5$. But now suppose we draw a different assignment vector $Z' = \{0, 1, 0, 1, 0, 1\}$. In which case, the ATE estimate becomes $10 - 6.7 = 3.3$. Z and Z' are equally likely assignment vectors given equiprobable treatment assignment. Neither perfectly estimates the true causal effect for this population of six individuals. Different potential assignments therefore result in different ATE estimates even though the sample (and potential outcomes) are constant. Hence, for any single random draw of assignments there will be uncertainty over the estimate as a result of the randomisation procedure.

Variance in the estimated ATE for different assignment schedules is explained by treatment effect heterogeneity. In our simplified case, one of the six individuals is impervious to the treatment. The effect of treatment on this individual is zero, while for the other five individuals the effect is a uniform increase of 5. Causal uncertainty in the *average* treatment effect is a function of treatment effect heterogeneity – a common, and increasingly important, concern in experimental research (Grimmer, Messing and Westwood, 2017). When a given treatment has heterogeneous effects on a given sample, then the specific random assignment of treatment will result in an unbiased albeit uncertain estimate of the true ATE.

In general, when we extrapolate from a finite set of observations to a broader phenomenon, there will always be some form of uncertainty. This can be as a result of a finite number of units sampled from a wider population, or it can be as a result of inherently stochastic procedures like randomisation. Our statistical estimation in both cases are a best effort to infer, from incomplete data, a generalisable claim. Causal uncertainty is frequently ignored by experimentalists, despite its natural connection to randomisation procedures and our growing focus on heterogeneity in treatment effects. In fact, claims about the nature of the uncertainty estimated by researchers are rarely made explicit in research. Despite this, causal uncertainty is important to experimentalists – the controlled randomisation of a prime, vignette, or other condition means there are unrealised potential assignments over the estimate as a function of how treatment was randomised – irrespective of the number of observations relative to the population.

Moreover, understanding what type of uncertainty one is estimating has a substantive impact on the values in question. As Abadie et al. (2014) show, conventional estimates of uncertainty (based around the assumption of sampling uncertainty) can be overly conservative estimates of the causal uncertainty – namely, the uncertainty due to other potential random treatment assignments. And as other statistical features are added to the data, like clustering or multiple observations per individual, researchers want to avoid erroneous claims about this uncertainty due to assumptions in the variance estimator about either the assignment or sampling procedures.

Sometimes experimentalists may simply want to ensure that their intervention worked – that is, for those sampled, to be confident that the observed change in outcome is a result of the treatment rather than a fluke of randomisation. In which case, experimenters are not worried about sampling uncertainty at all. The settings may be so far removed from "natural" populations that the question is not whether others' would react in the same way, but whether the actual instruments are effective at all.[4]

On the other hand, experimenters often want to estimate how likely the observed effect holds across a wider population. This is most often the case in randomised survey experi-

[4] Note also that sampling uncertainty is not the same as external validity – it is not a claim about how likely any observed effect is "in the wild". In the case where a researcher is concerned about whether the effect would hold given a different sample, the validity of that claim is still only with respect to the identical experimental conditions as it was originally conducted – visiting a lab or online portal, reading the same vignettes, exposure to the same rules and so forth.

ments where participants are sampled from much larger populations like general electorates. In which case, the more frequent concern about sampling uncertainty is important and should be embedded in the variance estimation strategy. Crucially, this concern is separate from the "causal" uncertainty in the parameter estimate.

In summary, experimentalists should choose their variance estimator dependent on the type(s) of uncertainty they want to account for. Two separate types of uncertainty – causal and sampling uncertainty – are particularly relevant to experimentalists. They also help structure how we appreciate extensions to variance estimation in the presence of grouped, or clustered data. The remainder of this paper focusses on this specific issue. How should experimentalists select their variance estimator, and justify its use, based on the experimental design? In the next section, I present principled guidance on variance estimation in experimental contexts where there are group-level features to the data.

2 Group-level structure in experimental data

Motivating examples

Study 1. Consider an observational study that assesses how a specific teaching resource correlates with student learning outcomes. Researchers collect student-level educational attainment measures across multiple classrooms, and record whether each classroom is using the resource or not. Note in this observational context access to the teaching resource is not controlled by the researcher. To estimate a correlation with students' performance, the researchers regress individual students' educational attainment on this binary indicator:

$$Y_i = \beta \times R_{C_i} + \epsilon_i, \tag{3}$$

where β is the effect of the teaching resource, C_i is the classroom which individual i belongs, and R_{c_i} is an indicator variable for whether i's classroom uses the teaching resource or not.

Crucially while data are sampled at the student-level, there are unobserved effects invariant within students of the same classroom but which vary across classroom assignments. For example, the qualities of individual classroom teachers (their enthusiasm, experience, teaching style) will be shared across students from the same class, but will vary across students in different classrooms. Since these factors will plausibly affect the outcome, but are unobserved in the data, the residuals of our statistical estimation will exhibit a group-level structure. More specifically, if characteristics of the specific classrooms play any role in students' educational attainment, then some part of the residual error from the classroom size regressor will be correlated by individual. That is, we can decompose the residual term such that $\epsilon_i = \epsilon_{C_i} + \gamma_i$, where ϵ_{C_i} is the error associated with unobserved effects at the classroom level, and γ_i is the random error component centred around 0.[5]

Study 2. Now consider a different study where researchers randomly assign whether classrooms use the teaching resource or not. Researchers again measure the same educational performance metric at the student level (Y_i). Each individual's exposure to the resource treatment is determined by their membership of a specific classroom. As in Study 1, therefore, this predictor is invariant within classrooms. Unlike in Study 1, however, whether a *classroom* has the resource is determined by random assignment. To estimate the effect of the teaching resource, the researchers estimate a very similar model to that in Study 1:

$$Y_i = \beta \times \mathrm{D}_{C_i} + \epsilon_i, \tag{4}$$

where β is the effect of the treatment, C_i is the classroom to which i belongs, and D_{C_i} is the randomised treatment assignment of classroom C_i. As in Study 1, the same unobserved classroom-level effects like teacher quality may separately impact the outcome. So again it seems reasonable to disaggregate ϵ_i into separate classroom-level and random components.

[5] I shall assume in all the following examples that there are no relevant covariates to control for. This assumption is unrealistic, but the focus of this paper is on potential bias of the treatment regressor's variance, and not in the variance estimates of covariates. Therefore we can ignore these covariates for the sake of simplicity.

Study 3. Finally, consider a different experiment where students take part in repeated rounds of an experimental game. Researchers randomly manipulate a binary feature of the game (the treatment) each round. Since there are multiple rounds students will, in expectation, be exposed to both treatment and control conditions. Data is collected at the individual-round level. Thus researchers estimate the following model:

$$Y_i = \beta \times D_{it} + \epsilon_{it}, \tag{5}$$

where i indexes individual students, t indexes rounds of the experiment, D_{it} is the round-specific treatment assignment of i and ϵ is the round-level error.

Let us assume that there are no learning or other order effects at play in this scenario. D_{it} is randomised at the individual round-level such that our estimate of the size of β will be unbiased both within and across individuals. But since we have multiple observations for each individual, unobserved individual-level characteristics will be invariant across portions of our dataset. For instance, suppose that the experimental study measures some form of cognitive-performance like multiplying two numbers together. The treatment involves some additional distraction during the task. Regardless of the presence of this distraction, it is likely that some portion of the performance for each individual will be driven by their (unobserved) pre-treatment cognitive ability. Since this ability is invariant within individuals, and individuals complete multiple rounds across both treatment conditions, the residuals for each individual ϵ_{it} will likely be correlated. Again, therefore, our intuition is that the error term of the model can be decomposed into a group level component (in this case the student) and a random component.

Despite their different designs, all three studies have very similar data structures as summarised in Table 2. In each case the data exhibits clear clustering – by classroom in Studies 1 and 2, and by individual in Study 3. In Studies 1 and 2 note that the group members are students within classrooms, whereas in Study 3 the units of analysis are repeat observations taken from the same students. Each study has a single a binary predictor,

Table 2: Comparison of the data structures for three hypothetical studies.

Study 1		*Study 2*		*Study 3*	
Observation	R	Observation	D	Observation	D
$Y_{c_1,i=1}$	1	$Y_{c_1,i=1}$	1	$Y_{i_1,t=1}$	1
$Y_{c_1,i=2}$	1	$Y_{c_1,i=2}$	1	$Y_{i_1,t=2}$	0
$Y_{c_2,i=3}$	0	$Y_{c_2,i=3}$	0	$Y_{i_2,t=1}$	0
$Y_{c_2,i=4}$	0	$Y_{c_2,i=4}$	0	$Y_{i_2,t=2}$	1
...		...		...	
$Y_{c_C,i=I-1}$	0	$Y_{c_C,i=I-1}$	0	$Y_{i_G,t=1}$	0
$Y_{c_C,i=I}$	0	$Y_{c_C,i=I}$	0	$Y_{i_G,t=2}$	1

R is an observational indicator about whether the main predictor is present or not. D is similarly binary indicator, except it is randomly assigned at the group-level (Study 2) or at the observation level (Study 3).

multiple observations per group, and an expectation that unobserved group-level factors will impact the outcome.

How should we handle the potential impact of group-level factors on the error term of our predictor in each study? In terms of observational studies like Study 1, the guidance in the statistical literature is clear: cluster your standard errors (Angrist and Pischke, 2009).[6] Otherwise, your estimates of the uncertainty around your point estimates will be biased.

Should the same logic apply to either of the two experimental studies? On the one hand, Study 2 and Study 3 both have similar data structures to Study 1 and share the propensity for group-level influences on the outcome. Unlike Study 1, however, the predictor is randomised in both Study 2 and Study 3.[7] Does the fact that the predictor is randomised preclude the need to cluster?

Studies 2 and 3, moreover, differ from each other in terms of the level at which treatment is randomised. In Study 2, treatment is assigned at the group level. All observations within a group share the same treatment assignment. In Study 3, on the other hand, treatment is

[6] Indeed, the classroom example is a modification of the motivating example in Angrist and Pischke's (2009) textbook example of when cluster-robust estimation is useful.

[7] Note, in Study 2, the randomisation across classrooms should ensure that features like teacher quality are balanced across treatment and control, even though observations are drawn at the student-level. While the parameter estimate will be unbiased, therefore, group-level features will likely impact the error terms as discussed above, and do so in a way that is correlated by classroom membership.

randomly assigned at the unit of observation and, consequently, we do not expect treatment assignment to be correlated within groups. More formally, let ρ_x denote the relationship between treatment assignment for observations within groups. ρ_x is the intraclass correlation coefficient of treatment assignment. In Study 2, we expect $\rho_x = 1$ since treatment assignment within each group is perfectly correlated. In Study 3, $\rho_x = 0$ in expectation. Does this difference mean that clustering is appropriate in one case but not the other?

Conflicting advice for clustering experimental errors

The literature is not settled on the answer to these questions (e.g. see Blattman, 2015), and specific and detailed guidance for experimentalists remains limited. Indeed, two recent working papers on the correct application of clustered variance estimators come to seemingly conflicting conclusions about contexts with randomised predictors (Cameron and Miller, 2015; Abadie et al., 2017). Whereas Cameron and Miller (2015) argue that randomisation typically precludes the need to cluster one's variance estimates, Abadie et al. (2017) suggest clustering is necessary in some, but not all, experimental protocols.[8]

These arguments are seemingly incompatible because they implicitly consider different types of design. Neither paper is solely focussed on contexts with randomised predictors, and as a result their prescriptions do not adequately generalise to the diversity of experimental designs common across the social sciences. In the remainder of this section I highlight the key insights from this research to provide a comprehensive guide to variance estimation in experimental contexts.

Cameron and Miller (2015) argue that 'if we think that either the regressors or the errors are likely to be uncorrelated within a potential group, then there is no *need* to cluster within that group' (p.17, emphasis added). It is unclear, however, what makes clustering unnecessary from their discussion. On the one hand, this could be interpreted as suggesting

[8] Another tempting solution is to check whether clustering makes a difference and report whichever error is larger. As discussed in the introduction, however, this conservative approach is misguided to the extent it conflates hypothesis testing with variance estimation, and misconstrues the fundamental statistical properties of the variance estimate.

that when a regressor is randomised there will be no difference in cluster-robust and conventional variance estimates (and therefore clustering is not needed). This claim can be tested using Monte Carlo simulations. In Section 3 I demonstrate that in fact randomisation does not preclude differences in variance estimates. Therefore, this strategy does not give us a principled reason to choose one variance estimator over another.

On the other hand, a more robust interpretation is that one *should* not cluster because in expectation the correlation will be zero in randomised contexts. In contexts where treatment is randomised at the individual-level, any non-zero correlation is a fluke of a random assignment procedure rather than bias caused by group-level clustering into treatment or control. This argument is more promising – it does provide a principled approach to variance estimation – but only holds if treatment is uncorrelated with groups in expectation. For instance, in Study 3 treatment is randomised at the observation-level, so not only is the point estimate itself unconfounded but *in expectation* the within-group correlation of treatment (ρ_x) should be zero. If, as in Study 2, treatment is randomly assigned across groups but is correlated within groups, then by definition ρ_x is non-zero. Therefore, cluster-robust and conventional estimators may differ. Which estimator is correct in this case, and why? And what if, aside from randomisation, researchers have sampling-related concerns about the groups in the data?

Design-based approach to clustering

Abadie et al. (2017) provide a general solution that incorporates similar intuitions to Cameron and Miller (2015) but which can be extended to clustered treatment assignment designs common in experimental research. Their approach builds on the broader theory behind experimental uncertainty discussed in Section 1. When choosing the correct variance estimator in the presence of clustered data and randomisation, experimentalists should consider both how clustering affects the sampling and treatment assignment of observations respectively. In short, clustering is appropriate, they argue, if either the assignment of treatment or sampling are conducted at the group-level.

Group-level treatment assignment. Consider first the case of a completely randomised experiment like Study 3. The estimated variance of the effect is, by definition, unrelated to the treatment. Even if there are group-correlated errors, the intergroup correlation of treatment assignment is zero in expectation and therefore we should not expect any bias in the variance estimate (under the causal uncertainty interpretation).[9]

With a single random treatment assignment schedule, however, there is a non-zero probability that $\rho_x > 0$ even with completely randomised treatment. For instance, in Study 3 any given subject may get treated more often than not (or vice versa) precisely because assignment is random.[10] The observed correlations in treatment at the group-level are a quirk of an (as-if) random procedure conducted at the observation-level. They do not reflect inherent correlations in treatment assignment within groups. If we were to repeat treatment assignment multiple times, the correlations would average to zero. It is therefore not appropriate to cluster on the basis of these correlations because they exist purely by chance.[11] Note this claim is made without reference to the data. We can know whether $\rho_x = 0$ in expectation purely from the design of the treatment assignment.

When treatment is assigned at the group-level, however, error-correlations due to group-level unobserved factors will not be cancelled out (in expectation) by the randomisation of treatment across groups. In this context, the conventional variance estimator *is* inappropriate. Suppose you have two pre-existing groups – one is assigned to treatment and one

[9] This logic is very similar to what Cameron and Miller (2015) appear to imply. The difference is that Abadie et al. (2017) make clear that the rationale behind clustering is not made with reference to any observed treatment correlation in the data.

[10] Consider flipping a fair coin ten times. In expectation we should get five heads and five tails but, inherent to the randomness of the procedure, we may plausibly get six, seven, or even eight heads from the ten coin flips. Randomness does not, in and of itself, guarantee balance empirically. Random procedures are random precisely because the assignment to one value or another is determined by some probabilistic mechanism where the functional form is both known and independent of all other variables. See Titiunik (2020) for a good discussion on the various definitions of 'random'.

[11] In a similar sense, researchers would not re-run an experiment just because a certain individual received the treatment condition in six out of ten rounds (when the expected proportion would be five out of ten rounds).

to control. Suppose further that some unobserved feature uniformly affects the outcomes of individuals in the treatment group. Since all observations in that group receive the same treatment assignment, there is no way for the estimation strategy to disentangle this group-level influence on the outcome (despite treatment assignment over the groups being random). Our errors should therefore account for these deliberate error correlations, and we do so by using the clustered variance estimator.

Group-correlated random treatment assignment shares features with observational studies where clustering is used to correct variance estimates for unobserved correlations between groups, predictors and outcomes. An unobserved contextual variable may influence outcomes independent of the effect of the main predictor. If groups of observations share this contextual feature as well as values for the main predictor, then conventional estimates of variance will not account for these group-level error correlations. In a similar way, if experimental treatments are correlated within groups, and groups share some background unobserved features that impact the outcome, then we need to correct for that in our estimation of the variance. This occurs in the hypothetical Study 2 because treatment is constant within classrooms and students within classes will share unobserved features that influence the residuals of the statistical model. In short, one should only cluster one's errors *on the basis of treatment assignment* if treatment assignment is correlated within groups.

Group-level sampling. Researchers making claims about a broader population should ensure variance estimates account for the absence of those units not sampled in their data. In the case of clustered data, the variance estimator should take into account whether the number of observed *groups* is part of a wider population of groups. Imagine a completely randomised experiment that is conducted on two individuals. They complete a set number of rounds, the treatment is randomised at the round-level, and the experiment is well-powered. One of these individuals, however, refuses to engage in the experiment and no matter what treatment condition they receive they always report the same outcome. This experiment will have two likely features. First, the point-estimate of the treatment is likely to be close to zero. But moreover, since the number of observations per individual is large, we expect the conventional standard errors to be quite narrow.

Intuitively these errors do not reflect the fact that an obstinate subject is skewing the results. At first glance, the narrow confidence intervals suggest the point-estimate is very precise. The problem in this case is that the researcher wants to make a claim about the experimental effect on a wider population of individuals but by chance one of the subjects is seemingly skewing not only the point-estimate but also the uncertainty of that estimate. By swapping the obstinate individual for another well-behaved subject, we might expect the treatment effect to change. Therefore, the uncertainty around the original estimate should account for this variability.

In such situations, clustering is appropriate since it captures the fact that the variance estimate is contingent on which groups were sampled from a wider population of groups.[12] In the case of the obstinate subject, we would expect this to widen the confidence interval around the low treatment effect estimate, to capture the fact that the average treatment effect in the population will likely be higher than that recovered in this specific sample of individuals. In other words, when we sample at the group-level (selecting some groups but not others), our variance estimate should be adjusted to account for the fact that not all groups are included in the data. Note that in this case we cluster regardless of how treatment is assigned, to capture sampling rather than causal uncertainty related to which groups are observed. Again, however, once the sampling strategy is determined, whether the researcher should cluster standard errors is determined without reference to the observed data itself.

In addition to sampling being conducted at the group-level, two additional features must hold for the cluster-robust estimator to be unbiased. First, the number of clusters in the data needs to be reasonably large since the cluster-robust estimator developed by Liang and Zeger (LZ-estimator; 1986) operates asymptotically (Cameron and Miller, 2015).[13] Second,

[12] Where repeat observations are taken, individuals can constitute groups (as in Study 3).

[13] The LZ-estimator can be expressed as $V[\beta] = (X'X)\hat{B}_C(X'X)$ where $\hat{B}_C = \sum_{g=1}^{G} X_g'\hat{u}_g\hat{u}_g'X_g$ and

the number of *groups* in the data must be a small proportion of the groups in the population or, equivalently, the number of groups in the population must be assumed to tend to infinity. If they are not, then cluster robust variance estimates will be conservatively biased against the true variance (Cameron and Miller, 2015; Abadie et al., 2017).

This "vanishing proportion" feature of the typical cluster-robust variance estimator has substantial implications for experimental researchers. When designing their analyses, experimenters must decide whether they wish to make an inference about a population beyond the groups included in the data. There are many cases where this will likely be the case – repeat-round games using a sample of individuals are typically used to infer the size of an effect on some wider human population (be it geographical or socio-economic). If generalisability is desired, then the requirement that the number of groups in the data is small relative to the population of groups must hold.

If the number of groups is not a vanishingly small proportion of possible groups, or researchers simply want to estimate the uncertainty of causal effects with respect to the specific sample, then cluster-robust estimation should not be used. In these cases, the data does not contain a vanishingly small sample of the clusters in the population – instead it contains the "population" of clusters itself. Consequently, the data's composition would violate the asymptotic assumptions needed for the cluster-robust estimator to be unbiased.

Summary. Table 3 distils the design-based approach and provides a simple reference for experimental researchers to use. If either treatment is assigned at the group level or the groups in the data are a small sample of groups from a larger population then clustering is appropriate (Abadie et al., 2017). If neither hold, then one should not use cluster-robust variance estimation. Two implications of this section are worth restating. First, researchers can determine the appropriateness of cluster-robust variance estimation before collecting

$\hat{u}_g = y - X_g\hat{\beta}$ (Cameron and Miller, 2015, p.8). Hence it is clear that the estimator operates by averaging over the groups, and will therefore perform better when the number of groups in the data is large. The optimal number of clusters will vary dependent on the features of the clusters themselves (Carter, Schnepel and Steigerwald, 2017). Corrections to the cluster-robust estimator can be made to account for small absolute numbers of clusters in the sample.

Table 3: Should researchers cluster based on the design-based approach to variance estimation?

	Sampled groups small proportion of all groups	*Sampled groups large proportion/all potential groups*
Treatment assignment at group-level	Cluster errors	Cluster errors
Treatment assignment at observation-level	Cluster errors (if making a population-level claim)	Do not cluster errors

data. Both the assignment and sampling strategies are known prior to data collection. Second, researchers should be sensitive to whether their analysis aims to make an inference about a population beyond their sample, and if so whether there are sufficiently many groups both in their sample and the wider population for the cluster-robust estimator to be unbiased. In all cases, researchers should make explicit the type(s) of uncertainty they quantify in their analysis.

What if experimenters choose the wrong estimation strategy? How big a difference does the choice of variance estimator make in practice, and therefore how biased can the standard errors be? In the next two sections I present simulation and replication evidence to show that the difference in variance estimation can be very substantial in experimental contexts. In turn, these results demonstrate the substantive importance of following the design-based approach outlined above.

3 Simulation evidence in favour of the design-based approach

This section expands upon the general problem of variance estimation in statistical terms. I focus in particular on the treatment assignment aspect of variance estimation. Sampling-related concerns are ubiquitous and covered in depth elsewhere (e.g. Huber, 1967; White, 1980; Liang and Zeger, 1986). Using Monte Carlo simulations, I demonstrate that even under observation-level randomisation it is probable that conventional and cluster-robust variance estimates will diverge substantially.

Conventional estimation of variance

Suppose some treatment D is linearly related to an outcome Y. The true model is specified as follows,

$$Y = D\beta + \epsilon, \tag{6}$$

where D is a $n \times k$ matrix, β is a $k \times 1$ vector, and ϵ is the randomly distributed residual. For the simplest experimental set up we consider, D can be considered a $n \times 1$ matrix of treatment assignments, and β the corresponding effect of treatment on the outcome.

With a finite sample of observations, the optimisation problem is finding $\hat{\beta}$ such that the squared residual $(\epsilon'\epsilon)$ is minimised. In Appendix A1, I present the full derivation for completeness. Here it is sufficient to state that our estimate of variance $var(\hat{\beta})$ is the squared difference between the true average causal effect (β) and our best estimate $(\hat{\beta})$, such that:

$$Var(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'], \tag{7}$$

which is equivalent to:

$$Var(\hat{\beta}) = (D'D)^{-1}E[\epsilon\epsilon']D(D'D)^{-1}. \tag{8}$$

The variance of $\hat{\beta}$ in Equation 17 is itself an estimate–in exactly the same sense that $\hat{\beta}$ is an estimate of the true β.[14] This feature is often neglected. Given a finite sample, we do not know the true variance of the treatment effect (and covariates). Instead, we observe a subset of the population, note the deviations from the predicted value (given the linear

[14] The Gauss-Markov theorem, which rests in part on the above assumptions, proves that $\hat{\beta}$ is the best linear estimator of β.

model above), and from this infer something about the true variance in the population. As such, the standard errors of the coefficients are themselves an estimate, and not a deductive by-product, of regression models.

Since X is known, from Equation 17 our estimate of variance reduces to $E[\epsilon\epsilon']$, which is the variance-covariance matrix (Ω). Under conventional estimation, we assume that $\Omega = \sigma^2 \times I$. This specification is a direct result of the standard Gauss-Markov (GM) assumptions. First, for every value of D, the conditional expectation $E[\epsilon\epsilon'|D]$ is assumed to be constant (i.e. variance is homoskedastic). Therefore, the diagonal elements of the variance-covariance matrix are equal to some constant σ^2. Second, since we also assume error terms are uncorrelated, there is no covariance between any two distinct residuals (i.e. off-diagonal elements of the covariance matrix are zero). Hence the variance-covariance matrix is equivalent to multiplying the $n \times n$ identity matrix by σ^2. Imposing these restrictions on the structure of Ω, we only need to recover an estimate of σ^2, using the observed error in our sample: $\hat{\sigma}^2 = \frac{e'e}{n-k}$.

Any estimate of the true variance of $\hat{\beta}$ is entirely dependent on the simplifying assumptions we make about the nature of the variance-covariance matrix Ω. In particular, if either of the two GM assumptions are violated then our estimate of the variance of $\hat{\beta}$ will be incorrect (by definition). This latter feature matters hugely for statistical inference. Consider the parallel concern about the point-estimates $\hat{\beta}$. When we violate the GM assumptions – if for instance observations are not independent – our estimated coefficients are no longer unbiased linear estimators of the true β. In which case, our statistical inferences using conventional estimation are invalid. The same logic applies to variance estimation. If we do not satisfy the GM assumptions, our estimation of the variance of $\hat{\beta}$ are incorrect, and therefore any statistical inference about β is harmed because we have incorrectly estimated its variance.

Biased standard errors under the assumption of clustered data

Clustering in the data is a specific violation of the GM assumptions, whereby the covariance between off-diagonal elements of Ω are non-zero. This is true if there is group structure to residuals, such that $cov(\epsilon_i, \epsilon_j) > 0$ if $g_i = g_j$, else $cov(\epsilon_i, \epsilon_j) = 0$. In which case, $\Omega \neq \sigma^2 I$ (since multiplying the identity matrix means all off-diagonal elements are equal to zero) and so σ^2 is not an unbiased estimate of $\hat{\beta}$'s variance.

When Ω exhibits clustering, we cannot reduce $E[\epsilon\epsilon']$ to an estimation problem over σ^2 alone. Instead, assuming homoskedastic residuals, the correct specification of the variance of β is

$$\hat{\text{var}}(\hat{\beta}) = \sigma^2(D'D)^{-1}D'VD(D'D)^{-1}, \tag{9}$$

where V is the block-diagonal matrix of inter-group error correlations ρ (Moulton, 1986). In other words, we have substituted the identity matrix, which imposes strict conditions on the values of the variance-covariance matrix, with a matrix of the same dimension as I but which is capable of capturing the extent to which residuals may covary when observations belong to the same group.

V itself is unobservable since we cannot observe the entire population and therefore do not know the actual covariation between groups. Valid inference about coefficients' variance when data has group-level structures requires an estimate of both σ^2 and V. Incorrectly using the naive estimator of variance, when residuals exhibit within-group covariance, leads to estimation bias (Moulton, 1986; Liang and Zeger, 1986). The extent to which the conventional variance estimator is biased (expressed as the ratio between the true and naive variance estimate of $\hat{\beta}$) can be expressed as,

$$\text{Bias}_{\text{var}(\hat{\beta})} = 1 + [(\text{var}(m_i)/\bar{m}) + \bar{m} - 1]\rho_x\rho, \tag{10}$$

where ρ_x is the intraclass correlation coefficient (ICC) of x_{gi}. The extent to which variance is inflated or deflated by is known as the "Moulton Factor" (Angrist and Pischke, 2009) after Brent Moulton who first outlined this relationship (Moulton, 1986).

As either the ICC or correlation of errors within groups increase (decrease), the standard error is inflated (deflated). Moreover, since both ρ and ρ_x have a possible range $[-1, 1]$ (in Moulton's specification) it is possible the conventional estimator can be both overly-conservative or not conservative enough of the true variance of the predictor.[15] As the mean group size $\bar{m}$ decreases, or observations become more unevenly distributed across groups, then this would also inflate the conventional estimate of the predictor's variance.[16]

Clustering under complete randomised assignment

The Moulton Factor captures the extent to which the conventional estimator is biased if, in fact, one should cluster standard errors in a given design. More broadly the Moulton Factor is a scalar that reflects the extent to which a variance estimate changes as one switches between using conventional and clustered variance estimators. Usefully, therefore, we can use the Moulton Factor to understand whether and why clustering alters variance estimation under the assumption of randomisation.

Let $g \in G$ index groups of observations that each contain j observations (for the sake of simplicity).[17] D_{ig} is the treatment assignment for the ith observation within group g and y_{ig} denotes the corresponding observed outcome. We estimate the model $y^*_{ig} = \beta_0 + \hat{\beta}_1 D_i g + \epsilon_{ig}$ and focus on the potential bias of $var(\hat{\beta}_1)$ which is given by Equation 10. Under these assumptions, the Moulton factor collapses to,

$$\text{Bias}_{\text{var}(\hat{\beta}_1)} = 1 + [j-1]\rho_x \rho. \tag{11}$$

[15] This range is true of the ICC estimator as detailed in Moulton (1986) (see Griffin and Gonzalez, 1995). More recent ICC specifications have a strictly positive range: given a random effects model $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$ for individuals i and groups j, the ICC can be expressed as $\frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\epsilon}$, i.e. the proportion of variance accounted for by the group over the total variance (Rodríguez and Elo, 2003). But note that even under the assumption that $\rho_x > 0$, if $\rho < 0$ then $\rho \cdot \rho_x < 0$.

[16] In a non-experimental setting, it is clear how a group-level regressor repeated in an individual-level dataset skews standard errors. Where the regressor is invariant within groups but variant across groups, $\rho_x = 1$ and so even with some small error-correlation within groups, the conventional estimator will be biased.

[17] In fact, what follows below is not contingent on equal group sizes. Unequal group sizes affect the size of the variance bias factor, but not the two correlation terms ρ_x and ρ. To the extent that one's dataset may contain missing data or uneven group sizes, the actual impact of group-related errors is mediated by the square-bracketed term in Equation 10, instead of that in Equation 11.

Hence, differences in the estimated variance rely on both the ICC and correlation of within-group residuals being $\neq 0$ and, trivially, that the number of observations per group must be > 1. Less trivially, as the number of observations per group increases, the potential bias of the conventional variance estimator increases (assuming some non-zero $\rho_x\rho$). Since both ρ and ρ_x must be non-zero for the estimated variance to be biased, I consider each inflation term in turn.

Correlated errors. Despite randomisation, it is likely that errors will be correlated within groups. Observation-level assignment will randomise treatment across observations within groups but that does not mean that inherent characteristics of that group are also randomised across observations. Suppose we run a repeat-observation experiment on a set of individuals with a binary treatment and some outcome like giving in a dictator game. The treatment may affect one's willingness to distribute, but each individual will have some unobserved qualities that affect their latent attitudes towards giving – altruistic preferences for example. These features do not affect the estimate of the treatment coefficient, but they may affect the estimate of the coefficient's variance. The same logic applies in contexts where randomisation occurs at the group-level. Unobserved features of the group will be invariant across these observations (and invariant with the treatment indicator too).

Since unobserved qualities are constant across rounds, irrelevant of treatment, there may be some constant unobserved effect of each group's profile. In which case, $\rho \neq 0$.

Intraclass correlation of the treatment regressor. $\rho \neq 0$ alone is insufficient to lead to a difference between conventional and cluster-robust variance estimates. The total inflation factor is a product of both the error correlation term and the intraclass correlation of the treatment regressor (i.e. $\rho \times \rho_x$). If $\rho_x = 0$, then irrespective of any correlation in error terms across groups, there will be no effect on the estimated variance of the treatment effect. This feature appears to motivate the claim in Cameron and Miller (2015) that randomisation precludes the need for clustered variance estimation.

In a fully randomised experimental context the expectation is that $\rho_x = 0$ since treatment is randomised within groups. Since treatment is randomly assigned at the observation level, $P(D_{ig}) \perp P(D_{ig'}), \forall g \neq g^{'}$. Hence the ICC of the treatment regressor will equal 0 in expectation. That is $\rho_x = 0$ when $P(D_{ig} = z) = P(D_{ig} = z'), \forall z, z' \in Z$ since the value of the regressor is not contingent on any feature of the group g, and so the proportion of variance explained by the grouping variable will be equal to zero.

In short, therefore, our estimates of randomised treatment effects *should not* be affected by repeat observations so long as the treatment is randomised within and across groups. Even if there is some group-correlated unobserved features in the data that are related to the outcome, when $\rho_x = 0$ the scaling factor collapses to 1 and the clustered and non-clustered errors are identical.

Despite this expectation, given the stochastic nature of random assignment, any particular randomisation may result in $|\rho_x| > 0$. Even a small observed ρ_x has the potential to bias one's variance estimate if either the group size or within-group error correlation is sufficiently large. In fact, the probability of observing a non-zero ρ_x is dependent on the number of rounds. When j is small the probability of achieving truly randomised allocations of treatments decreases and thus the likelihood that $\rho_x \neq 0$ increases.[18]

Under randomisation for a binary treatment,

$$\lim_{j \to \infty} P(D_{ig} = x) = 0.5, x \in \{0, 1\}.$$

But when the number of observations per individual j is relatively small, the within-individual proportion of treated and non-treated rounds may not be perfectly balanced. In a two-round experiment, for example, the probability of a completely imbalanced treatment profile is the same as a completely balanced profile: $P(D_1 = D_2) = P(D_1 \neq D_2) = 0.5$. And across the full sample, even if $P(D_{ij} = 1) = 0.5$, there may be sufficient imbalance within

[18] Though this will be somewhat offset by reducing the potential bias of the standard variance estimator relative to larger group sizes (see Equation 11).

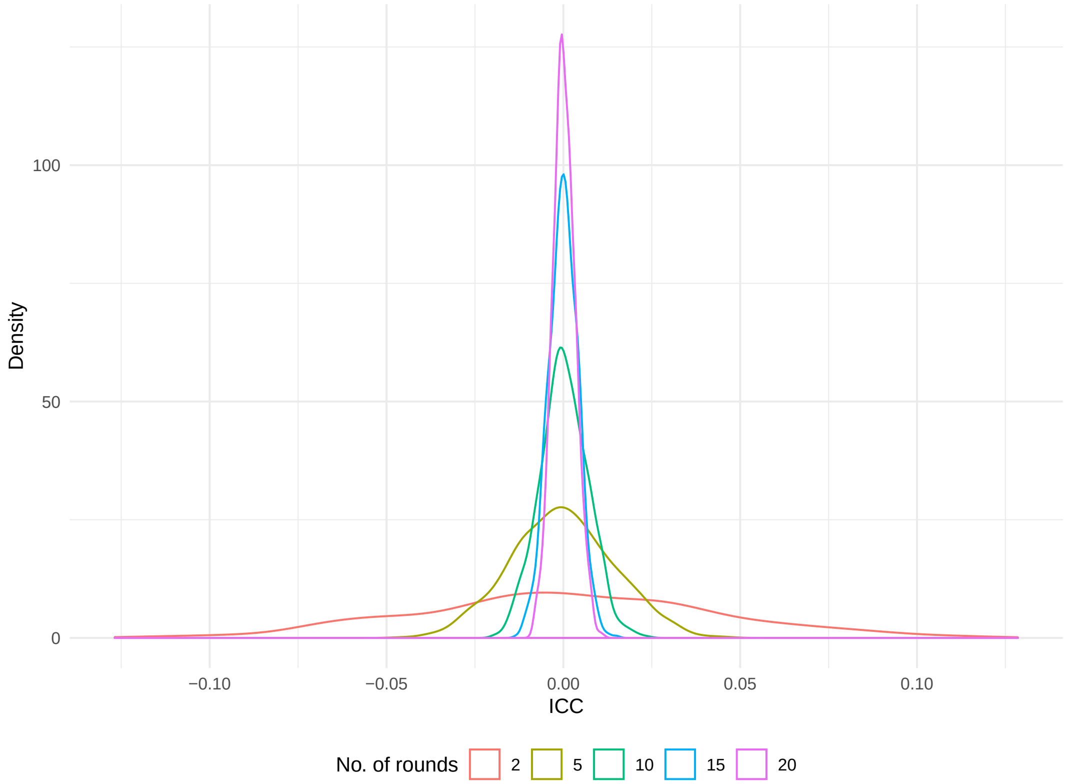

Figure 1: Distribution of Monte Carlo simulated intra-class correlations of the treatment regressor (ρ_x) with fully randomised treatment assignment over units. The coloured lines correspond to different numbers of observations per group j. Each simulation randomly assigns a binary treatment to $i \times j$ units, where $i = 500$. The distribution of values for each $j \in \{2, 5, 10, 15, 20\}$ is obtained by repeating the randomisation procedure 1000 times, calculating the ICC separately each iteration.

individuals' own set of assignments such that $\rho_x \neq 0$.

Figure 1 demonstrates this result using Monte Carlo simulations of ρ_x where treatment assignment is completely randomly assigned, for varying numbers of rounds per individual. The expected value across all round sizes is equal to zero, but with fewer rounds, there is clear potential for the observed ICC value of the regressor to be non-zero. At the lower bound of group sizes ($j = 2$) the ICC distribution is notably shallow but as group size increases to five, a 95 percent interval covers [-0.03,0.03].

Given the results in Figure 1, imbalanced treatment assignment within individuals as a result of random assignment will not always preclude within-group correlation of treatment. Thus, while the procedure of binary randomisation may itself be a stochastic process, there is no guarantee that the resultant treatment assignments *within* groups approximates a random distribution, such that there is some potential for $\rho_x \neq 0$.

In the case of group-correlated treatment assignment, $\rho_x \neq 0$ by definition. Therefore any group-level error structure as a result of unobserved features shared across members of a group will alter the estimated variance as determined by Equation 10.

Total change in variance. Non-zero values for both ρ and ρ_x are possible regardless of the level at which treatment is assigned. Multiple observations from the same group will likely mean errors are correlated within groups. Even when randomisation is conducted at the observation-level, this does not preclude non-zero within-group correlations for any given random treatment assignment. Clustering will therefore likely make a difference to one's estimate of the regressor's standard error. But by how much?

In Equation 11 j features as both an inflater of the entire scaling factor, but also deflates the variance of ρ_x's random distribution (see Figure 1). When these two forces are combined, even though an increase in the number of rounds lowers the probability of a non-zero ρ_x, this is undone by the larger $[j-1]$ inflationary term that acts separately to ρ_x in the scaling formula. Figure 2 displays the results of multiplying the computed ICCs in Figure 1 by the corresponding $[j-1]$. As a result, even at relatively large number of rounds per subject, the variance of the random ICC distribution around the mean of 0 is considerable.

Second, given ρ has a potential range $[-1, 1]$, the total distortion of the standard error is moderated by the extent of the error correlation. At its maximum $\rho = 1$, errors are entirely correlated within individuals, suggesting that all residual variance in the outcome is explained by invariant features of the subjects themselves. As $|\rho| \rightarrow 0$, this mediates the entire scaling factor, and thus the potential divergence from the conventional estimate decreases.

The results in this section demonstrate that conventional and cluster-robust variance estimates can vary substantially. These simulations also offer direct evidence against the ar-

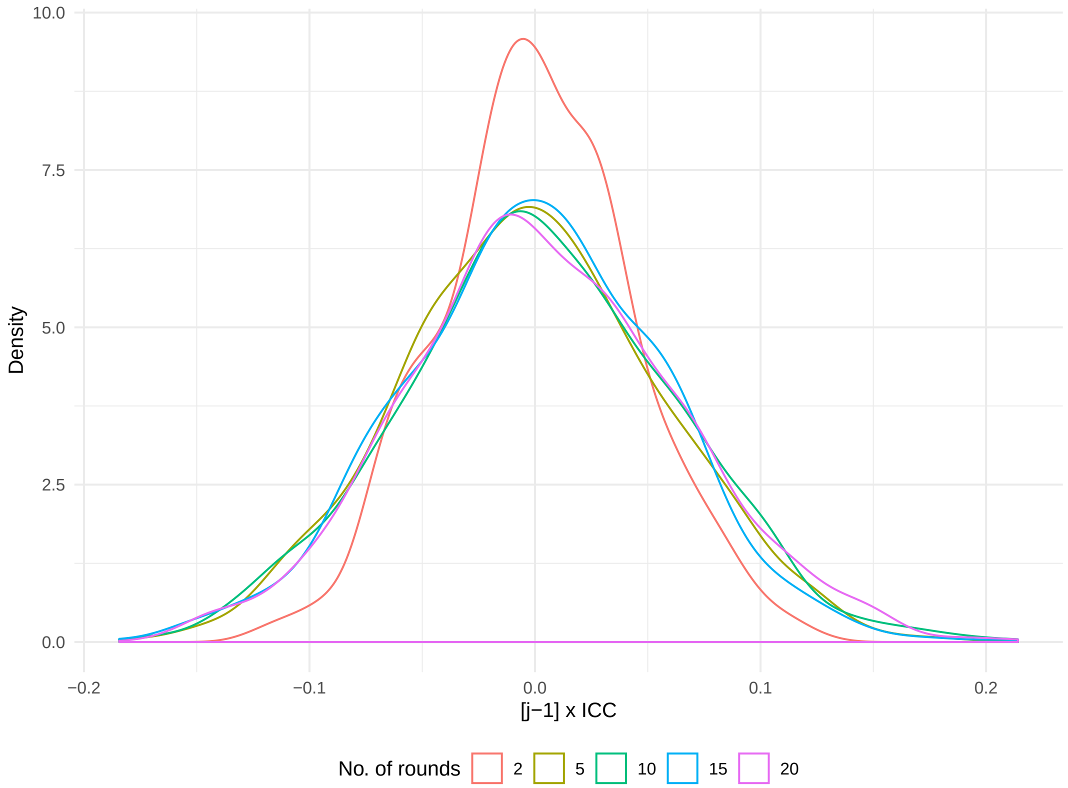

Figure 2: Distribution of the ICC estimates in Figure 1 multiplied by $[j-1]$. The effect of [j-1] is to offset any deflation in the variance of the ICC distribution as round size increases. Coloured lines again indicate different round sizes, holding fixed the number of individuals.

gument that randomisation precludes any difference between cluster-robust and conventional variance estimates. We are likely to observe differences in estimated standard errors between conventional and cluster-robust estimation, as a result of stochastic variation of treatment assignment within groups. These differences can be substantial, even for relatively simple experimental scenarios like the one explored in this section. Researchers should therefore pay careful attention to ensure they use the correct variance estimator, dependent on experimental design, as discussed in Section 2.

4 Review of Experimental Papers

To illustrate the importance of this paper's argument in substantive terms, this section presents a review of all papers published in the *Journal of Experimental Political Science*

(*JEPS*) between 2017 and 2019 (n = 49). *JEPS* publishes short (4000 word) studies using experimental methods, making it convenient to analyse the variance estimation methods used across a wide corpus of research involving randomisation.[19]

In total, all but two of the articles published between 2017 and 2019 in *JEPS* contain original experimental studies. For each article, I record the research domain, the variance estimation strategy, and any justification given for the method of variance estimation. I also note whether the articles satisfy either of the two clustering criteria given by Abadie et al. (2017) and the analysis in this paper. Table A1 in the Appendix presents a comprehensive list of all articles reviewed in this exercise.

Table 4 reports all articles that reported using clustered standard errors. In almost all cases, authors are correct to use the cluster-robust estimator. For instance, Zhang (2018) cluster their results by experimental session since treatment is assigned at the experimental session level (all individuals in one session receive the same treatment). A majority of the papers in Table 4 cluster on the basis of cluster-correlated treatments.

Notably, authors rarely correctly justify their use of these procedures (if they justify them at all). DeScioli and Kimbrough (2019), for example, note that "to control for repeated measures and group effects, we include random effects for participant, and we cluster standard errors at the group level." While random effects control for between subject differences, the repeated measures and group effects in this design alone are insufficient justification for clustering. Clustering makes sense both because the treatment assignment is correlated by group and because the individuals are plausibly drawn from a much larger population. In Demel et al. (2019), the authors note that "standard errors are adjusted parametrically to account for clustering within decision-making participants," which again is insufficient reasons to use the cluster-robust estimator without noting either the clustered sampling or assignment (both of which the authors' design includes).

[19] *JEPS* also states explicitly on its website its willingness to consider null results. Therefore, there is little *a priori* reason to suspect researchers would systematically modify their variance estimation in order to inflate the chance of recovering a statistically significant coefficient.

Table 4: *JEPS* articles (2017-19) using cluster-robust variance estimation

Authors	**Description**	**Sample**	**Cluster-level**	**N per cluster**	**No. of clusters**	**Clustered sampling?**	**Clustered treatment?**	**Should cluster?**
DeScioli and Kimbrough (2019)	Effects of information and communication on side-taking	*Not stated*	Group assignment	25	160	Yes	Yes	Yes
Demel et al. (2019)	Effects of employment status and ideology on redistributive preferences	Employed, unemployed, and students	Individuals	364	3	Yes	Yes	Yes
Loewen and MacKenzie (2019)	Effect of federalism on service representation	Political elites	Individuals	202	2	Yes	No	Yes
Zhang (2018)	Effect of enforcement institutions on likelihood of reporting corruption	Students	Session	20	18.5	No	Yes	Yes
Kingsley and Muise (2018)	Effect of communication in a public goods game	*Not stated*	Group assignment	21	25	Yes	Yes	Yes
Clayton (2018)	Effect of gender quotas on bias	Adults	Electoral district	135	7	Yes	Yes	Yes
Li and Zeng (2017)	Effect of facets of foreign direct investment on support in the recipient country	Adults	Individuals	1030	8	Yes	No	Yes
Klofstad (2017)	Effect of candidates' facial and vocal features on perceptions of competence	Adults	Candidate-pair	200	2	Yes	Yes	Yes
Green, Davenport and Hanson (2019)	Effects of Vietnam draft on behavour in the long term	Men	Birth year	3	300	No	No	**No**
Enos and Celaya (2018)	Effects of segregation on intergroup relations	Mturk/Student	Sessions	5	216	No	No	**No**

Two papers use cluster-robust variance estimation when it appears they should not. In Green, Davenport and Hanson (2019) the authors explore the long-term consequences of the Vietnam war draft on political behaviour. They leverage the natural randomisation of draft position being the result of lottery draws of birthdates within the year. The authors then cluster their standard errors by year since three separate random draws occurred between 1969 and 1971.[20] However, since the authors sample from across *all* draft lottery years, and the probability of assignment to "treatment" is equal between drafts, then neither the sampling nor treatment requirements of the design-based approach are satisfied.[21]

Similarly, in Enos and Celaya (2018), five separate experimental sessions are run on different subjects. In the pooled results reported in their Table 1, the results are clustered by session. Randomisation occurs *within* each experimental session. Moreover, it is unclear why or whether experimental sessions should be treated as sampling-based strategy. In the same sense that the draft lotteries constitute a population of potential clusters, the experimental sessions are the population of all sessions run. The potential "super-population" of all realised and unrealised experimental sessions seems, on the face of it, an unsatisfactory justification for clustering. The researchers did not, after all, sample from this set of potential groups prior to assigning treatment.

Among the set of papers that did not use cluster-robust variance estimation, four papers have designs where clustering would have been appropriate, and in two further studies a straightforward judgement is not possible. These papers are reported in Table 5.

Of the first four papers, half include cases where clusters were sampled from a wider population of clusters (Morton and Ou, 2019; Bassi, 2019; Barber et al., 2017). Morton and Ou (2019) note that clustering is appropriate for their design, on the basis of clustered sampling, but that privacy concerns precluded individuals from being identified in the data. Barber et al. (2017) similarly should cluster on the basis of clustered sampling. In this study,

[20] In an abstract sense, these draws can be though of as experimental sessions.

[21] The treatment is operationalised as having either a low or high draft number. Since lottery draws were exhaustive (all days in the year were drawn), the probability of assignment to a low draft number is approximately 0.5 in each year the lottery was conducted.

Table 5: *JEPS* articles (2017-19) with group-level structure that do not using cluster robust estimation

Authors	**Description**	**Sample**	**Clustered sampling?**	**Clustered treat-ment?**	**Should cluster?**
Morton and Ou (2019)	Effect of public voting on prosocial behaviour	Students	Yes	-	**Yes**
Bassi (2019)	Effect of weather on vote choice	Students & university employees	Yes	Yes	**Yes**
Lorenz, Paetzel and Tepe (2017)	Effect of framing taxes on redistribution preferences	*Not stated*	-	Yes	**Yes**
Jeon, Johnson and Robinson (2017)	Effect of ethnic diversity on social sanctioning	Two ethnic groups	-	Yes	**Yes**
Barber et al. (2017)	Effect of status quo bias in ballot wording on minority rights	Mturk	Yes	-	**Yes**
Rogers (2017)	Effect of risk elicitation on charitable giving	Students	Yes	-	Maybe

each subject considers multiple minority protections and is asked their opinion about the status quo. The researchers randomly vary whether a given minority right is presented as the status quo. Their analysis pools individuals' choices across the different minority protections. Since the unit of analysis is at the choice-level and individual subjects are plausibly a subset of a larger population clustering would have been appropriate.

The other two papers include treatments that are correlated at the cluster-level (Lorenz, Paetzel and Tepe, 2017; Jeon, Johnson and Robinson, 2017). In Lorenz, Paetzel and Tepe (2017), the authors conduct a multi-round group experiment in which participants decide on levels of redistribution. Two features are completely randomised per round: the endowment one receives, and the group to which you are assigned. On this basis, there is no clustered treatment assignment. However, the experimenters are interested in how framing redistributive choices affects individuals' choices. They therefore vary whether participants are told that the game is about taxation or a minimum income, and this framing occurs at the session-level. Individuals are assigned to only one session. Therefore, there are session-level groups where treatment is assigned. Thus, in this analysis, cluster-robust variance estimation is appropriate. Session-level clustering issues occur in Jeon, Johnson and Robinson (2017) study of ethnic diversity and social sanctioning. Individuals from two ethnic groups on the border of Kenya and Tanzania play a modified dictator game where a third-player can punish inequitable distributions by the dictator. The experimenters randomise the ethnicity of the different roles, leading to four separate treatment arms. As they note, when the ethnic composition of the session is balanced the four treatments are equally likely. But the composition of ethnicities within each session varies substantially (one session had a 71%-29% split). Since different ethnic compositions of sessions will affect treatment assignment probabilities, there is essentially clustered treatment assignment. The authors should therefore cluster at the session-level.[22]

Two studies proved more difficult to categorise. Rogers (2017) tests whether there are differences in risk elicitation experiments when individuals' earnings are given as charitable donations instead of direct payments to the participant. They use two separate risk elicitation experiments to test for these effects. Some, but not all, subjects play one of these games – the Bomb Risk Elicitation Task (BRET; Crosetto and Filippin, 2013) – twice. Therefore, in the analysis of this game there is a mixture of once- and twice-observed subjects. Later analyses focus just on those who completed BRET twice. The treatment over whether money is earned versus donated is orthogonal to assignment to the 2-round BRET game. Therefore there is no clustered treatment assignment in this experiment. However, since individuals are

[22] In this case the treatment probabilities are, in expectation, equal. To the extent this is a pre-data, design-related issue is debatable. If the randomisation procedure itself had produced imbalances purely by chance then clustering wouldn't be necessary. But since participation by members of ethnic groups is non-random, and the probabilities vary between sessions, it does seem pertinent to apply cluster-robust estimation.

a smaller sample of a large population and there are multiple rounds (for 40 participants) there is a cluster-based sampling component to the design. How one should handle this situation, given that the full analysis contains both single- and double-observations is not clear. One pragmatic solution would be to run the analyses entirely separately for the two groups, and cluster variance estimates for those observed twice in the BRET game.

Bassi (2019) explores the effect of weather on vote choice between candidates who are identical except for their "risk". Participants are randomly assigned to sessions held on days that had the largest difference in predicted precipitation, and complete ten rounds each of positively and negatively framed choices.[23] The payoff to choosing the risky candidate is sequentially altered over rounds such that rational voters should switch to voting for the non-risky option midway through each set of ten rounds. Bassi checks differences in preferences between positive and negative framing descriptively by comparing round-level proportions visually. To assess the statistical significance, however, Bassi sums the number of choices for the non-risky candidate per subject and conducts a Welch difference-in-means test. Similarly, to inspect the effect of weather, Bassi compares the total number of risky choices across all 20 rounds and compares again compares mean differences between good and bad weather.

This estimation strategy – summarising behaviour at the individual level – precludes clustering. However, it is not clear that this strategy is optimal. Some subjects may exhibit inconsistent behaviour, switching between candidates more than once. Rather than simply averaging away these issues, why not model the measurement error explicitly within the estimation strategy? A more robust strategy is to pool observations at the choice level, and model whether the risk-free incumbent is chosen on the basis of the difference in expected payoff, framing, and weather. In this proposed test, since the weather is invariant at the session-level cluster-robust variance estimation is appropriate on the basis of clustered treatment assignment. Bassi (2019) also implicitly draws individuals from a wider population,

[23] This framing is with respect to the earnings of a comparison country – either lower (positive) or higher (negative) than the expected payoff in the experiment.

and therefore clustering would similarly be appropriate on the basis of clustered sampling.

Replications

How does the reported variance change, and thus how are the results affected, when we correct the variance estimator used in the papers reported above? I re-estimate the two incorrectly clustered papers in Table 4 and provide revised analysis of the experimental results in Bassi (2019) from Table 5 using a pooled choice-level model.[24]

For Green, Davenport and Hanson (2019) and Enos and Celaya (2018), I compare an exact replication of their main results with the same estimation without using cluster-robust standard errors. Since Bassi (2019) does not pool results in the study itself, I construct a linear probability model for each of the four alternative weather measures used in the study.[26] I then compare the standard errors of the weather terms based on whether cluster-robust or conventional variance estimation is used.

Table 6 reports the results of this exercise. Each row is a separate model, and β refers to the main treatment coefficient. Overwhelmingly, the results show that correcting the variance estimation procedure increases the standard error of the treatment estimates – sometimes by a considerable amount. In both Green, Davenport and Hanson (2019) and Enos and Celaya (2018), conventional standard errors are *larger* than the original clustered standard errors. In neither paper does this change the substantive results,[27] but the percentage increase in the standard error is often over 10 percent and in the case of Enos and Celaya (2018) the standard error is almost doubled.

In Bassi (2019), the change in standard errors are even more stark. To be clear, the

[2425] Replication files are also available for Morton and Ou (2019) but, as noted previously, without subject identifiers it is not possible to compare clustered to non-clustered results.

[26] Model Specification:

$$P(\text{Vote for Incumbent})_{ij} = \alpha + \beta_1 \text{Difference in Payoff}_{ij} + \beta_2 \text{Framing}_{iij} + \beta_3 \text{Weather}_j + \epsilon_{ij},$$

where i indexes individuals and j experimental sessions.

[27] Indeed Green, Davenport and Hanson (2019) find null results for all estimations in their study. I focus on the partisanship measures, but the full paper considers other types of long-term effects too.

Table 6: Replication of *JEPS* articles with incorrect variance estimates

Coef.	N	β	SE	$\text{SE}_{\text{Cluster}}$	ΔSE	ΔSE (%)
Bassi (2019)						
Objective Weather	3320	-0.08	0.013***	0.033*	+0.02	+152.496
Subjective Weather	2660	-0.09	0.014***	0.032**	+0.017	+122.518
Absolute Rainfall	3320	0.06	0.013***	0.029	+0.016	+127.572
Relative Rainfall	3320	0.07	0.016***	0.025**	+0.009	+58.466
Enos and Celaya (2018): Original model clustered						
Black	1081	-0.12	0.032***	0.024***	+0.008	+34.014
White	1081	0.10	0.031**	0.016***	+0.016	+98.772
Green, Davenport, and Hanson (2019): Original model clustered						
Vote '12: Obama	899	0.82	3.339	2.915	+0.424	+14.5
Primary '16: Any Dem	648	-0.66	3.932	3.964	–0.032	–0.8
Obama Approval	898	-0.42	3.313	2.87	+0.443	+15.4
Party ID: Democrat	897	-3.82	3.221	3.032	+0.189	+6.2
Liberal / Very Liberal	912	-0.68	2.976	2.905	+0.071	+2.4
Don't Support Tea Party	907	1.03	2.525	2.232	+0.293	+13.1

author does not report pooled models and so this is not a direct replication. That said, the sign of each model is in line with Bassi's findings – good weather decreases the probability of voting for the safe candidate (holding the expected payoff constant). Rainfall (in absolute and relative terms) increases the likelihood of risk-averse choices. The difference between cluster-robust and conventional standard errors is very large. In three of the four models, the clustered standard error is over twice the size of the conventional estimate. Moreover, the absolute rainfall model coefficient becomes insignificant at conventional levels, which is notable given the respective means difference in Bassi (2019) is significant.

In summary, this section has applied the design-based criteria to assess how well published experiments accord with the design-based approach advocated in this paper and in Abadie et al. (2017). I find both cases where researchers incorrectly use clustered variance estimators and cases where cluster-robust estimation should have been used. In addition, for the three studies replicated in this section correcting the variance estimator has large inflationary effects on the reported variance – in one case, undermining a published finding. Irrespective of whether the results affect the degree of confidence in the (null) results, researchers should be concerned whether reported standard errors accurately reflect the uncertainty in their point estimates as an inherent estimation problem.

5 Practical guidance for experimentalists

Section 4 demonstrates that group-level data structures exist in a variety of experimental settings. Researchers should be clear about what type(s) of uncertainty they wish to capture and how this correlates with any group-level features of their data. If groups are sampled from a wider population of groups, or treatment is assigned at the group-level, cluster robust estimation is appropriate. Before concluding, it is worth briefly noting some concrete and practical implications for common types of experimental design.

Randomised controlled-trials. There are many reasons why researchers may assign treatments that are correlated at the group level. One particular instance is when treatment affects multiple observed units simultaneously (as in the hypothetical Study 2). This is most obviously the case in randomised controlled-trials (RCTs), which frequently involve group-level treatments (Peters, Langbein and Roberts, 2016; Kendall, 2003). For instance, Banerjee et al. (2007) conduct an RCT in which treatment is a standardized teaching programme assigned at the school year-group level. Maths and language scores at the student level are then compared across control and treatment schools for that year-group. Each student belongs to a school year-group in either treatment of control. Students of the same classroom, therefore, have perfectly correlated treatment assignment.

Clearly, in cases such as these where treatment is delivered at the group-level, $\rho_x \neq 0$ by design. All students in the same class receive the same treatment assignment, and therefore clustering standard errors is appropriate. Moreover, the magnitude of difference in standard errors between conventional and cluster-robust estimations will likely be much greater than in the fully-randomised context since ρ_x will equal 1 in expectation.

The researcher knows whether treatment assignment is assigned at the group-level prior to experimentation. How treatment is assigned is not a feature of the data, but of the experimental protocol. Researchers can ascertain whether clustering is appropriate on the basis of treatment assignment prior to estimation (and even pre-register it).

This type of experimental design has a lab-based parallel. In some lab experiments, treatment may be only feasible at the experimental-session level such that all participants for a given session receive the same treatment.[28] This might be the case because subjects interact in a common experimental game, or because some feature of the lab must be altered in order for the treatment to be delivered. In which case, as in the RCT design, $\rho_x \neq 0$ in expectation. Moreover, while experimental sessions are not naturally-occurring groups like villages (groups where we would expect unobserved group-invariant features that affect the error term), experimental session groupings are know to influence outcomes (see Fréchette (2012) for a full discussion of session-effects). In which case, not only *should* researchers cluster based on clustered assignment, but also we would expect the difference in variance estimations to be substantial.

Conjoint experiments. Conjoint experiments typify designs that have both grouped data and observation-level treatment (as in the hypothetical Study 3). Conjoint experiments involve presenting subjects with a combination of two profiles, and forcing subjects to choose which profile they prefer. The value for each attribute (the attribute level) for each profile is randomised (Hainmueller, Hopkins and Yamamoto, 2014). To increase the efficiency of these designs, experimenters typically ask subjects to make choices across multiple sets of

[28] Session-level treatment inherently captures some form of causal uncertainty. Given each group is assigned to one treatment arm or another, how different would our effect(s) be if the random treatment assignment differed?

candidates. For instance, in a two-profile design in which n subjects make k choices, the dataset is $n \times k \times 2$ rows long. Since the attributes are randomised at the profile-level, conjoint experiments do not satisfy the assignment-based justification for cluster robust estimation.[29]

Group-level error structures will arise in conjoint designs. Each $k \times 2$ subset of the data corresponds to a single individual. Therefore, any idiosyncratic features of the individual that affect their responses, or attitudes, will affect multiple observations in the dataset. For instance, while the potential values of each attribute are constant across individuals, how individuals interpret these values may differ (Dafoe, Zhang and Caughey, 2018). In which case, individual-level cognitive or semantic-processing differences may lead to error structures that are similar within an individual's set of responses, but differ between individuals.

The typical conjoint design does, however, employ cluster-based sampling. From the population of possible subjects, we randomly sample a vanishingly small number of individuals. Each of these individuals then completes multiple rounds of the experiment. Therefore, if we wish to make a generalised claim about the variance in the population, we *should* cluster standard errors. And since the population is likely to be much larger than the number of individuals sampled, we satisfy the asymptotic assumptions of the LZ estimator. If, however, our variance estimation is based on estimating the causal effect conditional on those chosen, then clustering is not required.

Researchers may wish to refrain from using cluster-robust variance estimators within conjoint designs if they either 1) want to assess the causal effect on this specific sample, or 2) field the experiment on a sample that likely contains a high proportion of clusters within the relevant population. Suppose we run a conjoint on politicians in a state legislature. Instead of having a small sample of subjects from a large population, with a sufficient response rate one's data would include a substantial proportion of all state legislators. In which case, it is inappropriate to cluster standard errors because the asymptotic assumption that $\frac{g}{G} \to 0$

[29] Unlike experimental designs discussed thus far, in the conjoint experiment there are multiple treatment coefficients. If each profile has A attributes and an average of L potential values per attribute, there will be $(A-1) \times L$ treatment coefficients (since each attribute has one omitted reference category). This does not substantively alter the variance explanation.

is not satisfied. This guidance holds even if there are idiosyncratic features of individual legislators that lead to correlated errors across each respondent's choices.

6 Discussion

In this paper, I present a comprehensive guide to cluster-robust estimation in experimental settings. I build on recent debates about variance estimation in the presence of randomisation (Blattman, 2015; Cameron and Miller, 2015; Abadie et al., 2017) and simplify the correct logic for applied researchers. In particular, I motivate the design-based logic developed by Abadie et al. (2017) and provide concrete estimation advice for applied experimentalists.

In short, if either treatment is assigned at the group-level or the groups in the data are a small sample from a wider population then clustering is appropriate. Conversely, experiments with large proportions of the group population and observation-level randomisation should not use cluster-robust variance estimators, irrespective of whether group-level features influence the outcome.

Using Monte Carlo simulations and replication data, I demonstrate that randomisation does not eliminate differences between conventional and cluster-robust standard errors. Researchers will likely observe differences between these two variance estimates, and have to make a substantive decision about which is the unbiased estimate. The principled guidance of this paper aims to encourage best practice, and debunk assumptions about randomisation's impact on variance estimates.

More generally, this paper aims to highlight that point-estimate uncertainty is itself an estimation problem. Given its fundamental importance to hypothesis testing, variance estimation should be treated as seriously as the corresponding point-estimates. Since the actual variation of the population is unknown, being aware of how the design of one's experiment influences the appropriateness of different variance estimators is key for valid inference about experimental treatment effects.

References

Abadie, Alberto, Susan Athey, Guido Imbens and Jeffrey Wooldridge. 2014. Finite Population Causal Standard Errors. Technical Report w20325 National Bureau of Economic Research Cambridge, MA: .

Abadie, Alberto, Susan Athey, Guido Imbens and Jeffrey Wooldridge. 2017. "When Should You Adjust Standard Errors for Clustering?" *arXiv:1710.02926 [econ, math, stat]* .

Abadie, Alberto, Susan Athey, Guido W. Imbens and Jeffrey M. Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88(1):265–296.

Amira, Karyn. 2018. "Do People Contrast and Assimilate Candidate Ideology? An Experimental Test of the Projection Hypothesis." *Journal of Experimental Political Science* 5(3):195–205.

Andersen, David J. and Richard R. Lau. 2018. "Pay Rates and Subject Performance in Social Science Experiments Using Crowdsourced Online Samples." *Journal of Experimental Political Science* 5(3):217–229.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton: Princeton University Press.

Arceneaux, Kevin. 2017. "Anxiety Reduces Empathy Toward Outgroup Members But Not Ingroup Members." *Journal of Experimental Political Science* 4(1):68–80.

Banerjee, Abhijit V, Shawn Cole, Esther Duflo and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *The Quarterly Journal of Economics* 122(3):1235–1264.

Barber, Michael, David Gordon, Ryan Hill and Joseph Price. 2017. "Status Quo Bias in Ballot Wording." *Journal of Experimental Political Science* 4(2):151–160.

Bassi, Anna. 2019. "Weather, Risk, and Voting: An Experimental Analysis of the Effect of Weather on Vote Choice." *Journal of Experimental Political Science* 6(1):17–32.

Bechtel, Michael M. and Kenneth F. Scheve. 2017. "Who Cooperates? Reciprocity and the Causal Effect of Expected Cooperation in Representative Samples." *Journal of Experimental Political Science* 4(3):206–228.

Blattman, Chris. 2015. "Clusterjerk, the Much Anticipated Sequel." https://chrisblattman.com/2015/12/11/clusterjerk-the-much-anticipated-sequel/.

Bokemper, Scott E., Peter Descioli and Reuben Kline. 2019. "Unfair Rules for Unequal Pay: Wage Discrimination and Procedural Justice." *Journal of Experimental Political Science* 6(3):180–191.

Bonilla, Tabitha and Cecilia Hyunjung Mo. 2018. “Bridging the Partisan Divide on Immigration Policy Attitudes through a Bipartisan Issue Area: The Case of Human Trafficking.” *Journal of Experimental Political Science* 5(2):107–120.

Buntaine, Mark T. and Lauren Prather. 2018. “Preferences for Domestic Action Over International Transfers in Global Climate Policy.” *Journal of Experimental Political Science* 5(2):73–87.

Busby, Ethan C. and James N. Druckman. 2018. “Football and Public Opinion: A Partial Replication and Extension.” *Journal of Experimental Political Science* 5(1):4–10.

Butler, Daniel M. and Charles Crabtree. 2017. “Moving Beyond Measurement: Adapting Audit Studies to Test Bias-Reducing Interventions.” *Journal of Experimental Political Science* 4(1):57–67.

Butler, Daniel M. and Miguel M. Pereira. 2018. “Are Donations to Charity an Effective Incentive for Public Officials?” *Journal of Experimental Political Science* 5(1):68–70.

Cameron, Colin A. and Douglas L. Miller. 2015. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources* 50(2):317–372.

Carter, Andrew V., Kevin T. Schnepel and Douglas G. Steigerwald. 2017. “Asymptotic Behavior of a t -Test Robust to Cluster Heterogeneity.” *The Review of Economics and Statistics* 99(4):698–709.

Clayton, Amanda. 2018. “Do Gender Quotas Really Reduce Bias? Evidence from a Policy Experiment in Southern Africa.” *Journal of Experimental Political Science* 5(3):182–194.

Costa, Mia. 2017. “How Responsive Are Political Elites? A Meta-Analysis of Experiments on Public Officials.” *Journal of Experimental Political Science* 4(3):241–254.

Crosetto, Paolo and Antonio Filippin. 2013. “The “Bomb” Risk Elicitation Task.” *Journal of Risk and Uncertainty* 47(1):31–65.

D’Adamo, Riccardo. 2019. “Cluster-Robust Standard Errors for Linear Regression Models with Many Controls.” *arXiv:1806.07314 [econ]* .

Dafoe, Allan, Baobao Zhang and Devin Caughey. 2018. “Information Equivalence in Survey Experiments.” *Political Analysis* 26(04):399–416.

Demel, Simona, Abigail Barr, Luis Miller and Paloma Ubeda. 2019. “Commitment to Political Ideology Is a Luxury Only Students Can Afford: A Distributive Justice Experiment.” *Journal of Experimental Political Science* 6(1):33–42.

Denny, Elaine K. and Jesse Driscoll. 2019. “Calling Mogadishu: How Reminders of Anarchy Bias Survey Participation.” *Journal of Experimental Political Science* 6(2):81–92.

DeScioli, Peter and Erik O. Kimbrough. 2019. “Alliance Formation in a Side-Taking Experiment.” *Journal of Experimental Political Science* 6(1):53–70.

Dietrich, Simone, Susan D. Hyde and Matthew S. Winters. 2019. "Overseas Credit Claiming and Domestic Support for Foreign Aid." *Journal of Experimental Political Science* 6(3):159–170.

Duch, Raymond, Denise Laroze, Thomas Robinson and Pablo Beramendi. 2020. "Multi-Modes for Detecting Experimental Measurement Error." *Political Analysis* 28(2):263–283.

Eicker, Friedhelm. 1967. "Limit Theorems for Regressions with Unequal and Dependent Errors." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1(Statistics):59–82.

Enos, Ryan D. and Christopher Celaya. 2018. "The Effect of Segregation on Intergroup Relations." *Journal of Experimental Political Science* 5(1):26–38.

Esarey, Justin and Andrew Menger. 2019. "Practical and Effective Approaches to Dealing With Clustered Data." *Political Science Research and Methods* 7(3):541–559.

Flynn, D. J. and Yanna Krupnikov. 2019. "Misinformation and the Justification of Socially Undesirable Preferences." *Journal of Experimental Political Science* 6(1):5–16.

Franco, Annie, Neil Malhotra, Gabor Simonovits and L. J. Zigerell. 2017. "Developing Standards for Post-Hoc Weighting in Population-Based Survey Experiments." *Journal of Experimental Political Science* 4(2):161–172.

Fréchette, Guillaume R. 2012. "Session-Effects in the Laboratory." *Experimental Economics* 15(3):485–498.

Green, Donald P. and Adam Zelizer. 2017. "How Much GOTV Mail Is Too Much? Results from a Large-Scale Field Experiment." *Journal of Experimental Political Science* 4(2):107–118.

Green, Donald P. and Lynn Vavreck. 2008. "Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches." *Political Analysis* 16(2):138–152.

Green, Donald P., Tiffany C. Davenport and Kolby Hanson. 2019. "Are There Long-Term Effects of the Vietnam Draft on Political Attitudes or Behavior? Apparently Not." *Journal of Experimental Political Science* 6(2):71–80.

Griffin, Dale and Richard Gonzalez. 1995. "Correlational Analysis of Dyad-Level Data in the Exchangeable Case." *Psychological Bulletin* 118(3):430–439.

Grimmer, Justin, Solomon Messing and Sean J. Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25(4):413–434.

Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22(01):1–30.

Hou, Yue and Kai Quek. 2019. "Violence Exposure and Support for State Use of Force in a Non-Democracy." *Journal of Experimental Political Science* 6(2):120–130.

Huber, Peter J. 1967. "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1(Statistics):221–233.

Huddleston, R. Joseph. 2019. "Think Ahead: Cost Discounting and External Validity in Foreign Policy Survey Experiments." *Journal of Experimental Political Science* 6(2):108–119.

Ihme, Toni Alexander and Markus Tausendpfund. 2018. "Gender Differences in Political Knowledge: Bringing Situation Back In." *Journal of Experimental Political Science* 5(1):39–55.

Imbens, Guido W and Michal Kolesar. N.d. "Robust Standard Errors in Small Samples: Some Practical Advice." . Forthcoming.

Israel-Trummel, Mackenzie L. and Ariela Schachter. 2019. "Does Shared Social Disadvantage Cause Black–Latino Political Commonality?" *Journal of Experimental Political Science* 6(1):43–52.

Jackson, John E. 2019. "Corrected Standard Errors with Clustered Data." *Political Analysis* pp. 1–22.

Jeon, Sangick, Tim Johnson and Amanda Lea Robinson. 2017. "Nationalism and Social Sanctioning Across Ethnic Lines: Experimental Evidence from the Kenya–Tanzania Border." *Journal of Experimental Political Science* 4(1):1–20.

Kendall, J M. 2003. "Designing a Research Project: Randomised Controlled Trials and Their Principles." *Emergency Medicine Journal* 20(2):164–168.

Kingsley, David C. and Daniel Muise. 2018. "More Talk, Less Need for Monitoring: Communication and Deterrence in a Public Good Game." *Journal of Experimental Political Science* 5(2):88–106.

Klofstad, Casey A. 2017. "Looks and Sounds Like a Winner: Perceptions of Competence in Candidates' Faces and Voices Influences Vote Choice." *Journal of Experimental Political Science* 4(3):229–240.

Kobayashi, Tetsuro, Asako Miura and Kazunori Inamasu. 2017. "Media Priming Effect: A Preregistered Replication Experiment." *Journal of Experimental Political Science* 4(1):81–94.

Leeper, Thomas J. 2017. "How Does Treatment Self-Selection Affect Inferences About Political Communication?" *Journal of Experimental Political Science* 4(1):21–33.

Levine, Adam Seth and Reuben Kline. 2019. "Loss-Framed Arguments Can Stifle Political Activism." *Journal of Experimental Political Science* 6(3):171–179.

Li, Xiaojun and Ka Zeng. 2017. "Individual Preferences for FDI in Developing Countries: Experimental Evidence from China." *Journal of Experimental Political Science* 4(3):195–205.

Liang, Kung-Yee and Scott L Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73(1):13–22.

Linos, Katerina and Kimberly Twist. 2018. "Diverse Pre-Treatment Effects in Survey Experiments." *Journal of Experimental Political Science* 5(2):148–158.

Loewen, Peter John and Michael Kenneth MacKenzie. 2019. "Service Representation in a Federal System: A Field Experiment." *Journal of Experimental Political Science* 6(2):93–107.

Lorenz, Jan, Fabian Paetzel and Markus Tepe. 2017. "Just Don't Call It a Tax! Framing in an Experiment on Voting and Redistribution." *Journal of Experimental Political Science* 4(3):183–194.

Merolla, Jennifer L. and Elizabeth J. Zechmeister. 2018. "Threat and Information Acquisition: Evidence from an Eight Country Study." *Journal of Experimental Political Science* 5(3):167–181.

Morton, Rebecca B. and Kai Ou. 2019. "Public Voting and Prosocial Behavior." *Journal of Experimental Political Science* 6(3):141–158.

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics* 32(3):385–397.

Neyman, Jerzy. 1990. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5(4):465–480.

Pedersen, Rasmus T. 2017. "Politicians Appear More Competent When Using Numerical Rhetoric." *Journal of Experimental Political Science* 4(2):129–150.

Pérez, Efrén O. and Margit Tavits. 2019. "Language Heightens the Political Salience of Ethnic Divisions." *Journal of Experimental Political Science* 6(2):131–140.

Peters, Jörg, Jörg Langbein and Gareth Roberts. 2016. "Policy Evaluation, Randomized Controlled Trials, and External Validity—A Systematic Review." *Economics Letters* 147:51–54.

Peterson, Erik and Gabor Simonovits. 2017. "Costly Values: The Limited Benefits and Potential Costs of Targeted Policy Justifications." *Journal of Experimental Political Science* 4(2):95–106.

Rodríguez, Germán and Irma T. Elo. 2003. "Intra-Class Correlation in Random-Effects Models for Binary Data." *The Stata Journal* 3(1):32–46.

Rogers, Jonathan. 2017. "Nothing to Lose: Charitable Donations as Incentives in Risk Preference Measurement." *Journal of Experimental Political Science* 4(1):34–56.

Schochet, Peter Z. 2013. “Estimators for Clustered Education RCTs Using the Neyman Model for Causal Inference.” *Journal of Educational and Behavioral Statistics* 38(3):219–238.

Scotto, Thomas J., Jason Reifler, David Hudson and Jennifer vanHeerde-Hudson. 2017. “We Spend How Much? Misperceptions, Innumeracy, and Support for the Foreign Aid in the United States and Great Britain.” *Journal of Experimental Political Science* 4(2):119–128.

Severson, Alexander W. 2018. “Partisan Affiliation and the Evaluation of Non-Prototypical Candidates.” *Journal of Experimental Political Science* 5(2):121–147.

Soroka, Stuart, Matthew Wright, Richard Johnston, Jack Citrin, Keith Banting and Will Kymlicka. 2017. “Ethnoreligious Identity, Immigration, and Redistribution.” *Journal of Experimental Political Science* 4(3):173–182.

Titiunik, Rocio. 2020. “Natural Experiments.” *arXiv:2002.00202 [econ, stat]* .

Trump, Kris-Stella and Ariel White. 2018. “Does Inequality Beget Inequality? Experimental Tests of the Prediction That Inequality Increases System Justification Motivation.” *Journal of Experimental Political Science* 5(3):206–216.

White, Ariel, Anton Strezhnev, Christopher Lucas, Dominika Kruszewska and Connor Huff. 2018. “Investigator Characteristics and Respondent Behavior in Online Surveys.” *Journal of Experimental Political Science* 5(1):56–67.

White, Halbert. 1980. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica* 48(4):817.

Wooldridge, Jeffrey M. 2003. “Cluster-Sample Methods in Applied Econometrics.” *The American Economic Review* 93(2,):133–138.

Zhang, Nan. 2018. “Institutions, Norms, and Accountability: A Corruption Experiment with Northern and Southern Italians.” *Journal of Experimental Political Science* 5(1):11–25.

A Appendix

A.1 Uncertainty as a function of randomisation

In the main text, Table 1 presents a particular case in which there is uncertainty over the treatment effect due to randomisation – what Abadie et al. (2014) call "causal uncertainty". In short, these issues arise because of the many potential random treatment assignments only one is ever realised. Given we can only ever estimate the treatment effect (we do not observe the full outcomes schedule) our estimate may differ dependent on the specific treatment assignment vector drawn. There are other plausible configurations of potential outcomes that generate similar concerns about causal uncertainty.

This section briefly motivates some alternative cases that collectively highlight why experimentalists should care about this causal uncertainty. In what follows, assume that our data is the full relevant population, such that sampling concerns drop out.

The first configuration demonstrates a Type I error whereby an estimated effect ($\hat{\tau}$) is non-zero when the true causal effect (τ) is precisely zero. In which case, without an appropriate measure of uncertainty, a single randomisation will yield a false positive result. Table A1 presents a potential outcomes schedule and treatment assignment for which this state of affairs holds.

Table A1: $\tau = 0, \hat{\tau} > 0$

i	$Y_i(1)$	$Y_i(0)$	D_i
1	10	10	1
2	5	5	0
3	10	10	1
4	5	5	0

Given these potential outcomes τ is precisely zero – there is no effect on any subject within the population. However, given the random assignment D_i, the best estimate of the treatment effect is the difference in means between those in treatment and control i.e.

$10 - 5 = 5$. This estimate is, technically, unbiased. As we increase the N the estimate will tend towards zero. But with this particular random assignment vector, it would appear there is a positive and substantial causal affect. To prevent a false positive, our measure of uncertainty should account for this type of chance occurrence even as the true population parameter is zero and the potential outcomes remain fixed.

In the configuration above, while the causal effect (of 0) is constant across individuals, the control/treatment outcomes vary. This sort of scenario depicts a highly simplified case where individuals hold different preferences/outcomes of which none are affected by an experimental stimulus. The treatment itself is homogeneous (it is precisely zero) but heterogeneity in the potential outcomes across individuals creates some leeway for our unbiased estimator to return a non-zero effect estimate.

It is also possible to construct cases where causal uncertainty yields a Type II error. Table A2 presents an alternative potential outcomes schedule that yields the converse to Table A1.

Table A2: $\tau > 0, \hat{\tau} = 0$

i	$Y_i(1)$	$Y_i(0)$	D_i
1	5	10	1
2	5	10	1
3	5	0	0
4	5	0	0

In this case, there are clearly two separate effects of treatment. For individuals $i = 1$ and $i = 2$ the effect of treatment is negative, whereas for the other two individuals it is positive. This might be the case in some sort of experiment where an information prime corrects individuals prior perceptions of some phenomena or quantity. Since the absolute size of the effect is the same across these two subgroups, the true population *average* treatment effect is zero. Under D_i in Table A2, the estimated average treatment effect is positive.

As in the first potential outcomes schedule in Table A1, this configuration exhibits heterogeneity in the potential outcomes. Unlike in the first case, the heterogeneity is present both in the control condition and the direction of the treatment effect. The treatment itself has a heterogeneous effect. Individuals' outcomes under treatment are identical precisely because the treatment acts in opposing directions for the two subgroups with different control outcomes.

The example potential outcomes in the main text demonstrate another issue where the true population average treatment effect is smaller than the estimated average treatment effect but still non-zero given some non-compliance. Common to this example and the further two supplementary configurations in this appendix section, uncertainty in the estimated effect arises as a result of stochastic assignment to either treatment or control.

Of course, the extent to which all members of particular subgroups are assigned to the same treatment status diminishes with the number of observations (assuming a truly random assignment). Part of the attraction of experimental methods is that we expect balance across covariates precisely for this reason – with equiprobable treatment probabilities and a sufficiently large sample any feature that may confound the effect estimate should be equally distributed among treatment conditions.

That being said, a truly random procedure will not necessarily ensure perfect balance. If there is any heterogeneity to the treatment effect, or to outcomes among individuals, different equiprobable assignments will lead to different estimated average treatment effects. As researchers, we aim to making inferences about the true treatment effect. The confidence we place in our estimates should therefore capture this uncertainty as a function of randomisation, separate from any sampling-related concerns.

A.2 Derivation of variance

Suppose some treatment D is linearly related to an outcome y:[30]

$$Y = D\beta + \epsilon, \tag{12}$$

To mimimise the squared residuals, rearrange Equation 12 to define the residual as $\epsilon = Y - X\beta$, multiply this expression by its transpose, and expand to get the quantity of interest[31]:

$$\epsilon'\epsilon = y'y - 2\beta'D'y + \beta'D'D\beta. \tag{13}$$

To find $\hat{\beta}$, we differentiate with respect to β, set the first-order condition to zero and solve:

$$\begin{aligned}
\frac{d\epsilon'\epsilon}{d\beta} &= -2D'y + 2D'D\hat{\beta} \Rightarrow 0 \\
D'D\hat{\beta} &= D'y \\
(D'D)^{-1}D'D\hat{\beta} &= (D'D)^{-1}D'y \\
\hat{\beta} &= (D'D)^{-1}D'y \qquad (14)
\end{aligned}$$

For convenience later on, we can substitute Equation 12 back into Equation 14, to express the estimated coefficients in terms of the true coefficients and the residual:[32]

[30] This derivation is heavily based on course notes from by Michael J. Rosenfeld entitled "OLS in Matrix Form", which can be accessed at `https://web.stanford.edu/~mrosenfe/soc_meth_proj3/matrix_OLS_NYU_notes.pdf`.

[31] This step relies on the transposition of matrices. First as it applies to the expansion of the squared residuals:

$$\begin{aligned}
\epsilon'\epsilon &= (y - D\beta)'(y - D\beta) \\
&= (y' - \beta'D')(y - D\beta) \\
&= y'y - y'D\beta - \beta'D'y + \beta'D'D\beta,
\end{aligned}$$

and second as it applies to simplifying the expanded regression:

$$\begin{gathered}
y'D\beta \equiv \beta'D'y \\
\therefore y'y - y'D\beta - \beta'D'y + \beta'D'D\beta \equiv y'y - 2\beta'D'y + \beta'D'D\beta
\end{gathered}$$

[32] Here, note that $(AB)^{-1}(AB) = I$ and therefore $(X'X)^{-1}X'X$ drops out of the equation.

$$
\begin{aligned}
\hat{\beta} &= (D'D)^{-1}D'(D\beta + \epsilon) \\
&= (D'D)^{-1}D'D\beta + (D'D)^{-1}D'\epsilon \\
&= \beta + (D'D)^{-1}D'\epsilon && (15)
\end{aligned}
$$

Our estimate of variance $Var(\hat{\beta})$ is the squared difference between the "true" coefficient (β) and our best estimate ($\hat{\beta}$), such that:

$$Var(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \qquad (16)$$

Substituting $\hat{\beta}$ with Equation 15,

$$
\begin{aligned}
Var(\hat{\beta}) &= [((D'D)^{-1}D'\epsilon)((D'D)^{-1}D'\epsilon)'] \\
&= E[(D'D)^{-1}D'\epsilon)(\epsilon'D(D'D)^{-1})] \\
&= E[(D'D)^{-1}D'\epsilon\epsilon'D(D'D)^{-1})] \\
&= (D'D)^{-1}E[\epsilon\epsilon']D(D'D)^{-1} && (17)
\end{aligned}
$$

A.3 *JEPS* review

Table A3: All articles published in *JEPS* between 2017 and 2019

Authors	Type of experiment	Sample	Main result	Observation-level	Obs	Cluster?
Pérez and Tavits (2019)	Survey experiment (phone)	Bilingual adults	Table 1	Individual	262	No
Morton and Ou (2019)	Lab experiment	Students	Figure 1	Individual-period	140	No
Loewen and MacKenzie (2019)	Survey (elite)	Political elites	Table 2	Politician-contact	404	Yes
Levine and Kline (2019)	Field experiment	Women members of community group	Table 1	Individual	100708	No
Israel-Trummel and Schachter (2019)	Survey experiment	African American adults	Figure 1	Individual	1200	No
Huddleston (2019)	Online survey experiment	Mturk	Figure 2	Individual	1512	No
Hou and Quek (2019)	Survey experiment	Nationally representative	Table 1	Individual	824	No

Green, Davenport and Hanson (2019)	Natural/survey experiment	Men	Table 5	Individual	900	Yes
Flynn and Krupnikov (2019)	Survey experiment	Nationally representative	Table 2	Individual	1031	No
Dietrich, Hyde and Winters (2019)	Lab experiment	Students	Table 5	Individual	777	No
DeScioli and Kimbrough (2019)	Lab experiment	*Not stated*	Table 1	Individual-round	4000	Yes
Denny and Driscoll (2019)	Survey experiment	Nationally representative	Table 2	Individual	120	No
Demel et al. (2019)	Lab-in-the-field experiment	Employed, unemployed, and students	Table 1	Individual-allocation	1092	Yes
Bokemper, Descioli and Kline (2019)	Online experiment	Mturk	Table 2	Individual	542	No
Bassi (2019)	Lab experiment	Students/university employees	Table 5	Individual-round	1660	No
Zhang (2018)	Lab experiment	Students	Table 2	Individual	371	Yes
White et al. (2018)	Online experiment	Mturk	Figure 1	Individual	2006	No

Trump and White (2018)	Survey experiment	Nationally representative	Figure 1	Individual	1014	No
Severson (2018)	Survey experiment	Nationally representative	Figure 2	Individual	918	No
Merolla and Zechmeister (2018)	Survey experiment	Adult	Figure 1	Individual	Numerous	No
Linos and Twist (2018)	Survey experiment	Nationally representative	Table 1	Individual	1000	No
Kingsley and Muise (2018)	Lab experiment	*Not stated*	Table 2	Individual-round	1260	Yes
Ihme and Tausendpfund (2018)	Survey experiment	Students	Table 1	Individual	377	No
Enos and Celaya (2018)	Online Experiment	Mturk/Student	Table 1	Individual-round	1081	Yes
Clayton (2018)	Natural experiment	Adults	Table 1	Individual	996	Yes
Butler and Pereira (2018)	Survey experiment	Political elites	Table 1	Individual	463	No
Busby and Druckman (2018)	Survey experiment	Students	Table 1	Individual	384	No

Buntaine and Prather (2018)	Online experiment	Mturk	Figure 1	Individual	1140	No
Bonilla and Mo (2018)	Survey experiment	Nationally representative	Figure 1	Individual	514	No
Andersen and Lau (2018)	Survey experiment	Mturk	Table 1	Individual	364	No
Amira (2018)	Survey experiment	Mturk	Figure 2	Individual	593	No
Soroka et al. (2017)	Survey experiment	Adults	Table 1	Individual	3783	No
Scotto et al. (2017)	Online survey experiment	Adults	Table 1	Individual	1010	No
Rogers (2017)	Lab experiment	Students	Figures 1/2	Individual-round	109	No
Peterson and Simonovits (2017)	Survey experiment	Nationally representative	Table 2	Individual	1924	No
Pedersen (2017)	Online survey experiment	Adults	Figure 1	Individual	1210	No
Lorenz, Paetzel and Tepe (2017)	Lab experiment	*Not stated*	Table 1	Individual-round	480	No

Li and Zeng (2017)	Online survey experiment	Adults	Figure 1	Individual-round	8240	Yes
Leeper (2017)	Online survey experiment	Nationally representative	Table 1	Individual	879	No
Kobayashi, Miura and Inamasu (2017)	Lab experiment	Students	Table 1	Individual	104	No
Klofstad (2017)	Online experiment	Adults	Table 1	Candidate	400	Yes
Jeon, Johnson and Robinson (2017)	Lab-in-the-field experiment	Two ethnic groups	Table 2	Game	89	No
Green and Zelizer (2017)	Field experiment	Republican women	Table 2	Individual	58751	No
Franco et al. (2017)	NA	NA	NA	NA	NA	NA
Costa (2017)	NA	NA	NA	NA	NA	NA
Butler and Crabtree (2017)	Field experiment	Political elites	Table 1	Individual	11801	No
Bechtel and Scheve (2017)	Online survey experiment	Nationally representative	Figure 3	Individual	8497	No
Barber et al. (2017)	Online survey experiment	Mturk	Table 2	Individual-choice	5720	No

Arceneaux (2017)	Online survey experiment	Mturk (white adults)	Table A6	Individual choice	1654	No

References

Abadie, Alberto, Susan Athey, Guido Imbens and Jeffrey Wooldridge. 2014. Finite Population Causal Standard Errors. Technical Report w20325 National Bureau of Economic Research Cambridge, MA: .

Abadie, Alberto, Susan Athey, Guido Imbens and Jeffrey Wooldridge. 2017. "When Should You Adjust Standard Errors for Clustering?" *arXiv:1710.02926 [econ, math, stat]* .

Abadie, Alberto, Susan Athey, Guido W. Imbens and Jeffrey M. Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica* 88(1):265–296.

Amira, Karyn. 2018. "Do People Contrast and Assimilate Candidate Ideology? An Experimental Test of the Projection Hypothesis." *Journal of Experimental Political Science* 5(3):195–205.

Andersen, David J. and Richard R. Lau. 2018. "Pay Rates and Subject Performance in Social Science Experiments Using Crowdsourced Online Samples." *Journal of Experimental Political Science* 5(3):217–229.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Arceneaux, Kevin. 2017. "Anxiety Reduces Empathy Toward Outgroup Members But Not Ingroup Members." *Journal of Experimental Political Science* 4(1):68–80.

Banerjee, Abhijit V, Shawn Cole, Esther Duflo and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *The Quarterly Journal of Economics* 122(3):1235–1264.

Barber, Michael, David Gordon, Ryan Hill and Joseph Price. 2017. "Status Quo Bias in Ballot Wording." *Journal of Experimental Political Science* 4(2):151–160.

Bassi, Anna. 2019. "Weather, Risk, and Voting: An Experimental Analysis of the Effect of Weather on Vote Choice." *Journal of Experimental Political Science* 6(1):17–32.

Bechtel, Michael M. and Kenneth F. Scheve. 2017. "Who Cooperates? Reciprocity and the Causal Effect of Expected Cooperation in Representative Samples." *Journal of Experimental Political Science* 4(3):206–228.

Blattman, Chris. 2015. "Clusterjerk, the Much Anticipated Sequel." https://chrisblattman.com/2015/12/11/clusterjerk-the-much-anticipated-sequel/.

Bokemper, Scott E., Peter Descioli and Reuben Kline. 2019. "Unfair Rules for Unequal Pay: Wage Discrimination and Procedural Justice." *Journal of Experimental Political Science* 6(3):180–191.

Bonilla, Tabitha and Cecilia Hyunjung Mo. 2018. "Bridging the Partisan Divide on Immigration Policy Attitudes through a Bipartisan Issue Area: The Case of Human Trafficking." *Journal of Experimental Political Science* 5(2):107–120.

Buntaine, Mark T. and Lauren Prather. 2018. "Preferences for Domestic Action Over International Transfers in Global Climate Policy." *Journal of Experimental Political Science* 5(2):73–87.

Busby, Ethan C. and James N. Druckman. 2018. "Football and Public Opinion: A Partial Replication and Extension." *Journal of Experimental Political Science* 5(1):4–10.

Butler, Daniel M. and Charles Crabtree. 2017. "Moving Beyond Measurement: Adapting Audit Studies to Test Bias-Reducing Interventions." *Journal of Experimental Political Science* 4(1):57–67.

Butler, Daniel M. and Miguel M. Pereira. 2018. "Are Donations to Charity an Effective Incentive for Public Officials?" *Journal of Experimental Political Science* 5(1):68–70.

Cameron, Colin A. and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50(2):317–372.

Carter, Andrew V., Kevin T. Schnepel and Douglas G. Steigerwald. 2017. "Asymptotic Behavior of a t -Test Robust to Cluster Heterogeneity." *The Review of Economics and Statistics* 99(4):698–709.

Clayton, Amanda. 2018. "Do Gender Quotas Really Reduce Bias? Evidence from a Policy Experiment in Southern Africa." *Journal of Experimental Political Science* 5(3):182–194.

Costa, Mia. 2017. "How Responsive Are Political Elites? A Meta-Analysis of Experiments on Public Officials." *Journal of Experimental Political Science* 4(3):241–254.

Crosetto, Paolo and Antonio Filippin. 2013. "The "Bomb" Risk Elicitation Task." *Journal of Risk and Uncertainty* 47(1):31–65.

D'Adamo, Riccardo. 2019. "Cluster-Robust Standard Errors for Linear Regression Models with Many Controls." *arXiv:1806.07314 [econ]* .

Dafoe, Allan, Baobao Zhang and Devin Caughey. 2018. "Information Equivalence in Survey Experiments." *Political Analysis* 26(04):399–416.

Demel, Simona, Abigail Barr, Luis Miller and Paloma Ubeda. 2019. "Commitment to Political Ideology Is a Luxury Only Students Can Afford: A Distributive Justice Experiment." *Journal of Experimental Political Science* 6(1):33–42.

Denny, Elaine K. and Jesse Driscoll. 2019. "Calling Mogadishu: How Reminders of Anarchy Bias Survey Participation." *Journal of Experimental Political Science* 6(2):81–92.

DeScioli, Peter and Erik O. Kimbrough. 2019. "Alliance Formation in a Side-Taking Experiment." *Journal of Experimental Political Science* 6(1):53–70.

Dietrich, Simone, Susan D. Hyde and Matthew S. Winters. 2019. "Overseas Credit Claiming and Domestic Support for Foreign Aid." *Journal of Experimental Political Science* 6(3):159–170.

Duch, Raymond, Denise Laroze, Thomas Robinson and Pablo Beramendi. 2020. "Multi-Modes for Detecting Experimental Measurement Error." *Political Analysis* 28(2):263–283.

Eicker, Friedhelm. 1967. "Limit Theorems for Regressions with Unequal and Dependent Errors." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1(Statistics):59–82.

Enos, Ryan D. and Christopher Celaya. 2018. "The Effect of Segregation on Intergroup Relations." *Journal of Experimental Political Science* 5(1):26–38.

Esarey, Justin and Andrew Menger. 2019. "Practical and Effective Approaches to Dealing With Clustered Data." *Political Science Research and Methods* 7(3):541–559.

Flynn, D. J. and Yanna Krupnikov. 2019. "Misinformation and the Justification of Socially Undesirable Preferences." *Journal of Experimental Political Science* 6(1):5–16.

Franco, Annie, Neil Malhotra, Gabor Simonovits and L. J. Zigerell. 2017. "Developing Standards for Post-Hoc Weighting in Population-Based Survey Experiments." *Journal of Experimental Political Science* 4(2):161–172.

Fréchette, Guillaume R. 2012. "Session-Effects in the Laboratory." *Experimental Economics* 15(3):485–498.

Green, Donald P. and Adam Zelizer. 2017. "How Much GOTV Mail Is Too Much? Results from a Large-Scale Field Experiment." *Journal of Experimental Political Science* 4(2):107–118.

Green, Donald P. and Lynn Vavreck. 2008. "Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches." *Political Analysis* 16(2):138–152.

Green, Donald P., Tiffany C. Davenport and Kolby Hanson. 2019. "Are There Long-Term Effects of the Vietnam Draft on Political Attitudes or Behavior? Apparently Not." *Journal of Experimental Political Science* 6(2):71–80.

Griffin, Dale and Richard Gonzalez. 1995. "Correlational Analysis of Dyad-Level Data in the Exchangeable Case." *Psychological Bulletin* 118(3):430–439.

Grimmer, Justin, Solomon Messing and Sean J. Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25(4):413–434.

Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22(01):1–30.

Hou, Yue and Kai Quek. 2019. "Violence Exposure and Support for State Use of Force in a Non-Democracy." *Journal of Experimental Political Science* 6(2):120–130.

Huber, Peter J. 1967. "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1(Statistics):221–233.

Huddleston, R. Joseph. 2019. "Think Ahead: Cost Discounting and External Validity in Foreign Policy Survey Experiments." *Journal of Experimental Political Science* 6(2):108–119.

Ihme, Toni Alexander and Markus Tausendpfund. 2018. "Gender Differences in Political Knowledge: Bringing Situation Back In." *Journal of Experimental Political Science* 5(1):39–55.

Imbens, Guido W and Michal Kolesar. N.d. "Robust Standard Errors in Small Samples: Some Practical Advice." . Forthcoming.

Israel-Trummel, Mackenzie L. and Ariela Schachter. 2019. "Does Shared Social Disadvantage Cause Black–Latino Political Commonality?" *Journal of Experimental Political Science* 6(1):43–52.

Jackson, John E. 2019. "Corrected Standard Errors with Clustered Data." *Political Analysis* pp. 1–22.

Jeon, Sangick, Tim Johnson and Amanda Lea Robinson. 2017. "Nationalism and Social Sanctioning Across Ethnic Lines: Experimental Evidence from the Kenya–Tanzania Border." *Journal of Experimental Political Science* 4(1):1–20.

Kendall, J M. 2003. "Designing a Research Project: Randomised Controlled Trials and Their Principles." *Emergency Medicine Journal* 20(2):164–168.

Kingsley, David C. and Daniel Muise. 2018. "More Talk, Less Need for Monitoring: Communication and Deterrence in a Public Good Game." *Journal of Experimental Political Science* 5(2):88–106.

Klofstad, Casey A. 2017. "Looks and Sounds Like a Winner: Perceptions of Competence in Candidates' Faces and Voices Influences Vote Choice." *Journal of Experimental Political Science* 4(3):229–240.

Kobayashi, Tetsuro, Asako Miura and Kazunori Inamasu. 2017. "Media Priming Effect: A Preregistered Replication Experiment." *Journal of Experimental Political Science* 4(1):81–94.

Leeper, Thomas J. 2017. "How Does Treatment Self-Selection Affect Inferences About Political Communication?" *Journal of Experimental Political Science* 4(1):21–33.

Levine, Adam Seth and Reuben Kline. 2019. "Loss-Framed Arguments Can Stifle Political Activism." *Journal of Experimental Political Science* 6(3):171–179.

Li, Xiaojun and Ka Zeng. 2017. "Individual Preferences for FDI in Developing Countries: Experimental Evidence from China." *Journal of Experimental Political Science* 4(3):195–205.

Liang, Kung-Yee and Scott L Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73(1):13–22.

Linos, Katerina and Kimberly Twist. 2018. "Diverse Pre-Treatment Effects in Survey Experiments." *Journal of Experimental Political Science* 5(2):148–158.

Loewen, Peter John and Michael Kenneth MacKenzie. 2019. "Service Representation in a Federal System: A Field Experiment." *Journal of Experimental Political Science* 6(2):93–107.

Lorenz, Jan, Fabian Paetzel and Markus Tepe. 2017. "Just Don't Call It a Tax! Framing in an Experiment on Voting and Redistribution." *Journal of Experimental Political Science* 4(3):183–194.

Merolla, Jennifer L. and Elizabeth J. Zechmeister. 2018. "Threat and Information Acquisition: Evidence from an Eight Country Study." *Journal of Experimental Political Science* 5(3):167–181.

Morton, Rebecca B. and Kai Ou. 2019. "Public Voting and Prosocial Behavior." *Journal of Experimental Political Science* 6(3):141–158.

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics* 32(3):385–397.

Neyman, Jerzy. 1990. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5(4):465–480.

Pedersen, Rasmus T. 2017. "Politicians Appear More Competent When Using Numerical Rhetoric." *Journal of Experimental Political Science* 4(2):129–150.

Pérez, Efrén O. and Margit Tavits. 2019. "Language Heightens the Political Salience of Ethnic Divisions." *Journal of Experimental Political Science* 6(2):131–140.

Peters, Jörg, Jörg Langbein and Gareth Roberts. 2016. "Policy Evaluation, Randomized Controlled Trials, and External Validity—A Systematic Review." *Economics Letters* 147:51–54.

Peterson, Erik and Gabor Simonovits. 2017. "Costly Values: The Limited Benefits and Potential Costs of Targeted Policy Justifications." *Journal of Experimental Political Science* 4(2):95–106.

Rodríguez, Germán and Irma T. Elo. 2003. "Intra-Class Correlation in Random-Effects Models for Binary Data." *The Stata Journal* 3(1):32–46.

Rogers, Jonathan. 2017. "Nothing to Lose: Charitable Donations as Incentives in Risk Preference Measurement." *Journal of Experimental Political Science* 4(1):34–56.

Schochet, Peter Z. 2013. “Estimators for Clustered Education RCTs Using the Neyman Model for Causal Inference.” *Journal of Educational and Behavioral Statistics* 38(3):219–238.

Scotto, Thomas J., Jason Reifler, David Hudson and Jennifer vanHeerde-Hudson. 2017. “We Spend How Much? Misperceptions, Innumeracy, and Support for the Foreign Aid in the United States and Great Britain.” *Journal of Experimental Political Science* 4(2):119–128.

Severson, Alexander W. 2018. “Partisan Affiliation and the Evaluation of Non-Prototypical Candidates.” *Journal of Experimental Political Science* 5(2):121–147.

Soroka, Stuart, Matthew Wright, Richard Johnston, Jack Citrin, Keith Banting and Will Kymlicka. 2017. “Ethnoreligious Identity, Immigration, and Redistribution.” *Journal of Experimental Political Science* 4(3):173–182.

Titiunik, Rocio. 2020. “Natural Experiments.” *arXiv:2002.00202 [econ, stat]* .

Trump, Kris-Stella and Ariel White. 2018. “Does Inequality Beget Inequality? Experimental Tests of the Prediction That Inequality Increases System Justification Motivation.” *Journal of Experimental Political Science* 5(3):206–216.

White, Ariel, Anton Strezhnev, Christopher Lucas, Dominika Kruszewska and Connor Huff. 2018. “Investigator Characteristics and Respondent Behavior in Online Surveys.” *Journal of Experimental Political Science* 5(1):56–67.

White, Halbert. 1980. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica* 48(4):817.

Wooldridge, Jeffrey M. 2003. “Cluster-Sample Methods in Applied Econometrics.” *The American Economic Review* 93(2,):133–138.

Zhang, Nan. 2018. “Institutions, Norms, and Accountability: A Corruption Experiment with Northern and Southern Italians.” *Journal of Experimental Political Science* 5(1):11–25.