The paper has been carefully reviewed by an AE and two referees. The two referee reports can be downloaded from EJMS.

Both reviewers find that this paper contains potentially somewhat interesting results, but that the paper in its current form is not suitable for AoS. The two reports are quite detailed and emphasize different aspects of the weaknesses of the paper. Both recommended rejection as they found it unclear if the contributions would be sufficient even with a better presentation. The AE also recommended rejection. 1° I think they want more practical motivations.

All felt that the results are of high quality but the statistical motivation is unclear, i.e., what is the broader significance to the statistical community? The question of significance from
both referees is valid. The first referee, who is not an expert in this special topic, is an expert in general causal inference. If you can overcome this substantial obstacle in a revision you can choose to resubmit the paper. In this case please refer to the original submission and include a detailed response to the AE and reviewers. However please note that we at this time do not see how you will overcome this obstacle of needing to demonstrate broader significance. As a result, at this stage, no promises could be given as to the eventual outcome.

Practical motivations for our regimes

As its title suggests, this article presents Berry-Esseen type of bounds for design-based estimators in causal inference. The main practical motivation seems to be the 2^K factorial design, which is repeatedly used in the article to illustrate their results for more general designs. I am not in a good position to judge the novelty and theoretical contributions of this work: this is partly due to my lack of expertise in this particular area and partly due to the quality of writing. Below I will offer some comments on the significance and the writing of the article.

1 Significance

Understanding the properties of linear estimators in the randomization model is both an old and a new problem. It is old in the sense that this model has been around since the beginning of modern statistics and is at the core of causal inference. It is new in the sense that its importance in causal inference has only been emphasized by some authors since recently. So why is there such a big gap historically?

Personally, I think the reason is that investigations under the randomization model have largely failed to tell us that we should analyze our experiments differently. Nowadays, most trialists analyze their data using some version of the generalized linear mixed-effect models (or Cox models for survival outcome). On the other hand, most works in the randomization model (including the present paper) consider linear models/estimators and in my experience the suggested methods are generally not too different from the typical solutions. It is nice to give a justification of what people have been already doing using the more precise randomization model, but without new methodologies this will remain a purely theoretical endeavor.

This is why I think the most interesting part of this paper is the covariance estimator in the unreplicated design in Section 3.2. However, I was disappointed by the ad hoc grouping strategy. Different analysts may use different groups and obtain different results; it's not clear how such inconsistencies can be dealt with, and how grouping may affect the power of the analysis. Moreover, how is this different from just collapsing the treatment levels in the same group and then perform the usual analysis?

I am not trying to say that theory is not important. In particular, unifying theorems such as those presented in this paper could be quite reassuring for serious statisticians. But it is important to distinguish methodological and theoretical contributions.

1° Seems that there needs to be more practical motivation of the five regimes.

2° this guy also cares much about method novelty.

3° Indeed a good point. Don't have an answer for now. But let's see what we can say.

2 Writing

The writing quality of this article is generally quite poor and not at the level expected for the Annals. Many concepts are defined and conditions are introduced without enough motivation; sometimes the explanations come a page too late. The notation is extremely heavy and difficult to follow. The results are organized in a poor way with a lot of back-and-forth. In many occasions, the writing is not precise enough.

Here are some examples:

1. It was emphasized several times that the theoretical result is based on formulating the estimator "in terms of a linear permutational statistic". But it is never explained what a linear permutational statistic means.

2. Abstract: "However, the existing theory for this framework is far from complete especially when the number of treatment levels diverges and the group sizes vary a lot across treatment levels." It is not clear what "group sizes" precisely mean in my first reading. Instead, the authors could have said the "treatment group sizes" or "the size of the treatment arms".

3. Abstract: what does "diverging dimension of causal effects" mean? Do you mean diverging number of contrasts?

4. Introduction: A central assumption is that the treatment $\mathbf{Z}$ is independent of the potential outcomes. The authors mention that Neyman treated all potential outcomes as fixed (which would imply the independence), but it is not clear if this assumption is made in this article as well.

5. The vector $\mathbf{Z}$ is in bold but other vectors and matrices are not.

6. page 2: "Li and Ding (2017) ... studying the properties of the moment estimator for γ". This is not accurate, because the only estimator of γ considered here is linear. The authors may be referring to the sample variance $\hat{S}(q, q)$, but that is not an estimator for γ.

7. page 2: I am not sure if equation (3) should be accredited to Li and Ding (2017), as indicated by the writing.

8. Table 1 needs to be introduced before the opening paragraph of Section 1.2 talks about the first regime (R1). It is not clear until several pages

later than the results of this article can be used in all regimes in Table 1.

9. Non-uniform design (Definition 3) seems to include Nearly uniform design (Definition 2) as a special case. So the terminology may be inconsistent.

10. Example 1: "$\pm Q^{-1}$ entries" $\to$ "entries of $\pm Q^{-1}$".

11. Explanations of condition 1 is only given a page later (and I think they are not enough).

12. Page 8: "Third, the upper bound in (12) decreases at the rate of $(H/N)^{1/2}$ which can deal with (R1)(R4)". Need to elaborate.

13. Page 9: No explanation of the condition in (14) is given. This is not considered in Example 3.

14. Page 10: "Both $\hat{T}$ and T are related to T_0" but T_0 is a generic χ^2 random variable. Writing is very casual.

15. middle of page 11: Theorem 3(i) "reviews" $\to$ "reveals"?

16. Condition 4 is stated in words and is open to interpretation. How is it related to $M_N(q)$ and $S_N(q,q)$?

17. Very little is said about the proof of the theorems, which make it difficult to assess the novelty and technical difficulty of these results.

My recommendation is that the authors need to completely rewrite the whole article and communicate their most important results in a clear and transparent way for a resubmission to be considered.

Review: "Berry–Esseen Bounds For Design-Based Causal Inference With Possibly Diverging Treatment Levels and Varying Group Sizes"

1 Background

In this paper, authors work in the design-based potential outcome framework commonly referred to as the Neyman / Rubin model. There are N experimental units and Q categorical treatment which may be given to each unit. Each unit $i \in [n]$ has Q potential outcomes, denoted $Y_i(1) \dots Y_i(Q)$. The mean outcome vector is defined as $\bar{Y} = 1/N \cdot (\sum_{i=1}^n Y_i(1), \dots \sum_{i=1}^n Y_i(Q))$. In the design-based framework considered in this paper, the potential outcomes are deterministic and the only source of randomness is treatment assignment itself. The experimenter specifies a distribution on the treatment assignments, which are the n random variables $Z_1, \dots Z_n$ taking values in $\{1, 2, \dots Q\} \triangleq [Q]$. The experimenter observes the (random) outcomes $Y_1 \dots Y_n$ defined as

$$Y_i = \sum_{q=1}^{Q} Y_i(q) \cdot \mathbb{1}[Z_i = q] \ .$$

All analysis is performed under a completely randomized design, which is defined as follows: the experimenter specifies integers $N_1, \dots N_q$ such that $\sum_{q \in Q} N_q = N$ and the design is to sample a treatment assignments uniformly from the set of all treatment assignments such that exactly N_q units are assigned to treatment q for each $q \in [Q]$. In other words, that $\sum_{i=1}^n \mathbb{1}[Z_i = q] = N_q$ with probability 1 for all $q \in [Q]$.

The authors consider the following class of estimands: an experimenter specifies a linear transformation represented by the matrix H-by-Q matrix F and the goal is to estimate the linear transformation of the mean potential outcome vector,

$$\gamma = F\bar{Y} \ ,$$

which is a (possibly) multi-variate estimand. The most common example of such an estimand is the *average treatment effect*, which is the quantity

$$\gamma = \frac{1}{n} \sum_{i=1}^{n} Y_i(1) - Y_i(0) \ ,$$

which corresponds to a linear transformation with $H = 1$. Another example discussed (but not defined) in the paper is aggregated factorial effects.

2 Main Results

The majority of the literature has focused on analyzing experimental settings where each treatment group is a constant fraction of the overall population, i.e. the ratio N_q/N is bounded below by a constant as the sample size N grows. Of particular interest in this paper are different regimes, where N_q may grow much more slowly with the sample size, e.g. $N_q/N \to 0$ for some $q \in [Q]$ or even $N_q = 1$. Note that if $N_q = 1$ for all $q \in [Q]$, then this means that $Q = N$ so that the number of treatments is equal to (and thus growing with) the sample size.

The main results in the paper are a series of multivariate Berry-Esseen type bounds between a standardized estimator and a limiting distribution, either a standard normal or a χ^2. Roughly speaking, there are two types of Berry-Esseen bounds developed in this paper. The first type (Section 2) demonstrates a normal limiting behavior for linear functions of the estimator while the second type (Section 3) demonstrates a χ^2 limiting behavior for a standardized quadratic estimator. Both of these results may be used to establish validity of various hypothesis tests. An additional result is the development of a grouping-based covariance estimator for the setting where $N_q = 1$, which is shown to be conservative and to facilitate the construction of valid confidence sets for the estimand.

3 Weaknesses of the Paper

While the paper does contain interesting and timely results, it suffers from three major drawbacks: poor motivation, lacking technical discussion, and technical difficulties. I list these factors below and expand on them in more detail later in the review. I believe these factors must be more well addressed before the paper is ready for publication in Annals of Statistics.

1. **Point 1: Poor Motivation**: The setting investigated by the authors is not well-motivated which questions the relevance of the extensions.
2. **Point 2: Lack of Technical Discussion**: There are numerous occasions in the paper where the authors present an assumption without proper elaboration on the condition. This makes interpretation of the scope conditions, and thus relevance of, the results challenging.
3. **Point 3: Technical Difficulties**: I believe there are a number of minor technical errors in the paper. Taken at face value, they seem easily fixable so I refer to them as "technical difficulties".

1° same to R1 : more practical examples.

Point 1: Poor Motivation A large part of the motivation of the paper is to categorize completely randomized experimental designs into 5 categories, based on the behavior of the treatment group sizes (pages 3-4). Authors carefully detail differences between each of these regimes. But the reader is left wondering: "which of these regimes are relevant to statistical practice?" For example, the only setting where (R4) seems to be relevant is factorial designs with $K = \log_2 N$ so that $Q = N$, which is certainly interesting, but the only example that comes to mind. It is therefore difficult to imagine an experimental setting where (R5) holds, that is a mixture of (R1)-(R4) and (R4). Moreover, it is difficult to imagine a realistic setting (beyond the factorial designs with $K = \log_2 N$) where the experimenter chooses to use a severely non-uniform designs when a more uniform design is possible. The standard response to this critique is "the experimenter may not get to choose their experimental design beforehand" but even here, it is hard to imagine realistic scenarios where a severely non-uniform design (i.e. $N_q = 1$, a major consideration of this paper) is employed. It is not clear to this reader that the experimental regimes should even be split up based on treatment

group sizes alone. For example, the regimes should also include the "complexity" or "regularity" of the estimand that is to be estimated, as this seriously affects the results as well.

Point 2: Lack of Technical Discussion There are many points throughout the paper where there is a lack of technical discussion, but I will highlight only a few such points.

- **Trade-off Between Design Non-Uniformity and Estimand Regularity:** The most glaring to me is the inherent trade-off between non-uniformity in the design and the "complexity" or "regularity" of the estimand itself. For instance, in the setting of (R4) where the treatment groups are of size 1, there are many linear estimands which cannot be estimated: in fact, the typical average treatment effect cannot be estimated. Indeed, the authors require special conditions on the matrix representation of the estimand in order for their results to hold; however, there is very little discussion regarding *what estimands satisfy these conditions or their practical relevance.* Without a more thoughtful technical discussion, it seems almost like a bait-and-switch: authors provide "unifying" central limit theorems under extreme amounts of non-uniformity in the design, but neglect to discuss that these results require equally restrictive conditions on the estimands under consideration. To be clear: I'm perfectly happy with such a trade-off, but only if it is clearly presented as such.
- **Explanation of Condition 1:** Condition 1 places limitations on the estimand as well as the potential outcomes. However, there is no discussion on its interpretation, aside from what it can accomplish in a proof. One would expect that it would be conceptually cleaner to place separate conditions on the potential outcomes and the linear function separately; this joint condition is opaque and further discussion is warranted. Lemma 1 provides sufficient conditions under which Condition 1 holds, but they appear to standard assumptions in certain restricted cases; it is not clear, for example, concretely what *new* experimental settings satisfy Condition 1 but perhaps not the sufficient conditions in Lemma 1. Without such a discussion, the scope of the technical results is very difficult to determine.
- **Appearance of W_N in Theorem 4:** Theorem 4 requires that the normalized covariance estimator $W_N = V_{\hat{\gamma}}^{1/2} \mathrm{E}[\hat{V}_{\hat{\gamma}}] V_{\hat{\gamma}}^{1/2}$ has a limiting distribution $W_N \to W_\infty$. There should be a discussion around what this means and when one should expect it to occur. The third part of Theorem 4 describes an asymptotically normal test statistic which uses W_N; however, the relevance of such a test statistic is unclear because the experimenter cannot construct W_N, as it relies on the unknown expectation $\mathrm{E}[\hat{V}_{\hat{\gamma}}]$. There should be a technical discussion as to the use to this quantity W_N and whether, or how, the experimenter should obtain it.

3.1 Minor Technical Difficulties

There are several minor technical difficulties that make the paper more difficult to read. I suspect that they can all be addressed easily, so I do not think they greatly affect my review. However, I want to present them in case it is helpful constructive feedback.

- **Typo**: In Line (3), authors use the variable $\widehat{Y}_q$, but this refers to a scalar-valued random variable. Instead, I believe they mean to refer to $\widehat{Y}$ which is the vector-valued random variable.
- **Definitions 2 and 3**: There should be greater care given in Definitions 2 and 3 to be explicit about the asymptotic sequence. For example, Definition 2 states that a design is *nearly uniform* if there exists a positive integer $N_0 > 0$ and absolute constants $c' \leq c''$ such that $c' \leq N_q/N_0 \leq c''$ for all $q \in [Q]$. It's unclear whether N_0 grows with N or not. As written,

we need more clarification here.

it appears that N_0 is an integer, which does not grow with N. However, given the context of Definition 3, it seems that N_0 does need to be growing with N, so that the corresponding treatment sizes N_q are considered "large".

- **Proof Extensions**: There are several instances throughout the paper where authors write something along the lines "this theorem can be strengthened, but we admit technical details". For example, in Remark 1, after Condition 4, and after Theorem 8. These technical details are not easy to fill in to the causal reader, and so I suggest proving the result in the appendix and referring to this. → 5° more technical details in main paper?

I have also included several minor stylistic suggestions which do not affect my review, but which I would like to bring to the author's attention.

- Authors define the complete randomization design in Definition 1. However, this is not a very intuitive definition to the causal reader. I would recommend adding a more intuitive description below, perhaps something along the lines of what I wrote in the beginning of my review. ✓
- In Line (3) and throughout the paper, authors use $\text{Var}(A)$ to refer to the covariance of a vector-valued random variable A. In their notation, $\text{Var}(A)$ is a matrix. I would suggest using $\text{Cov}(A)$ instead as this cannot easily be mistaken for a scalar. ✓