Response to the comments of referee 2

This paper discusses the nature of inferences that are warranted for different types of experimental designs used in applied work in economics. The paper is essentially a non-technical overview of issues relating to appropriate inference in different experimental (non-observational) contexts. It draws together a wide range of arguments from the existing literature (although it is not always easy to tell which studies are the source of specific aspects of the different arguments). A common theme is that with frequentist statistics it is important to be clear about the nature of the reference set of repeated samples. Given this is a short paper, discussion of individual points is often relatively brief, so readers unfamiliar with the arguments will often be left wanting more detail or clarification. As an example, the discussion of internal and external validity on p. 4, while useful, is covered in much more detail in Athey and Imbens (2017). The paper concludes with a set of relevant issues to consider when contemplating inference in different experimental designs. Overall, the paper provides a useful overview of potential inferential pitfalls with experimental data, and the comments below are minor queries.

Minor points

- *p.2: The self-declared objective is to "address the question of statistical and scientific induction and, more particularly, the role of the p-value for making inferences beyond the confines of a particular experimental study". It is not entirely clear why there is a particular emphasis on the p-value (rather than, say, standard errors or confidence intervals), as the main points relate to the relevance or otherwise of frequentist-based statistical inference in general, of which p-values are only one aspect. The focus on p-values may just reflect a continuation of an anti-p-value theme evident in Hirschauer et al. (2018, 2019); Mayo (2018) presents an alternative viewpoint on the place of p-values in frequentist statistical inference that is not subject to many of the concerns about p-values expressed in the wider literature.*

While the paper's focus is on *p*-values, our fundamental argument is indeed more generally concerned with the applicability and meaning of statistical inference based on random error as expressed in the standard error of the randomization or the sampling distribution. We included an additional note stating that we put a particular focus on *p*-values because of their high prevalence in empirical research and the scientific criticisms regarding *p*-value-based statistical practices.

- *p.2, para 4: 'Statistical independence' is referred to in parentheses, but the link to the preceding sentence could be explained in a little more detail. (The same point applies at the top of p.3.) At least in the context of regression estimation of average treatment effects, random assignment of the treatment does not imply that the error term in the regression of the observed outcome on a treatment dummy is independent of the latter (Athey and Imbens, 2017).*

Sorry, the reference in the parentheses may have been misleadingly short. We made the reference more complete by changing it into "statistical independence of treatments," a brief term that is meant to condense the crucial feature that randomization balances known *and* unknown confounders across treatment groups.

- *p.3: Arguably, the balance between the advantages of between-subject and within-subject designs is overly stacked towards the former. Which is 'better' will likely depend on context. Gelman (2019), for example, argues that "within-person designs are generally the best option when studying within-person effects". He points out that the main disadvantage of a between-subject design is that it does not control for variation across subjects, which can be unduly large and so dominate.*

We included a reference to this statement by Gelman (2019) and additionally emphasized that the adequateness of the experimental design depends on the research context.

- *p.3, lines 4-7: Order effects can be mitigated, to some extent, by counterbalancing in a repeated measures design.*

We included a brief comment on the mitigation of order effects through counterbalancing in repeated measure designs.

- *p.3, last para: In the context of discussing a two-independent-sample t-test, it is noted that "the resulting p-value targets the following question: when there is no treatment-group difference, how likely is it that we would find a difference as large as (or larger than) the one observed when we repeatedly assigned the experimental subjects at random to the treatments under investigation". Shouldn't this statement refer to the calculated t-statistic value, not the size of the difference? The test statistic could be 'large' because of a small standard error rather than a large group difference. A similar query arises with the passage on p.4, para 3.*

For a given standard error derived from a single study, referring to the sample statistic (e.g. a difference) and referring to the test statistic (e.g. the *t*-value) in the interpretation of the *p*-value is equivalent. We believe, however, that referring to the sample statistic of interest (here: the group difference) is more intuitive. Vogt et al. (2014: 242), the reference we provide, and many others use similar wordings.

- *p.5: When discussing cases where the sample is the finite population, it would be worth referring to Abadie et al. (2014) who consider a meaningful role for standard errors in such contexts.*

We included a reference to Abadie et al. (2014) in the discussion of this issue.

Typos, etc.

- *p.2, line 7 (and several other instances, including the list of references): Athey and Imbens (add 's')*
- *p.3: Dunning (2013) in the text but 2012 in the reference list.*
- *p.4, fn.4: Wasserstein et al. (2016) is not in the reference list.*
- *p.8, line 8 up: 'Fourth' instead of 'Forth'*

Typos etc. are corrected.