# Assignment 3 – Analysis of article “Storm crowds: evidence from zooniverse on crowd contribution design” (Barbosu and Gans, 2017, NBER)

Within this study, the researchers investigate how a task design change within a crowd science platform affects the contribution quantity, completeness and quality of the contributions. The context used for this article is the “Zooniverse” crowd science platform, the largest of its kind worldwide, in which a particular project folded 2 previously separate tasks into a single one, which is termed as a reduction in contribution divisibility by the authors.

The research question that this research aims to address is how contribution level and quality is affected by the design of the crowdsourcing community. In particular, how the divisibility of contributions, i.e. whether minor contributions like correcting a typo is considered a contribution (high divisibility) or whether an original article counts as a contribution (low divisibility), factors into contribution behavior of platform members.

Predictions based on a formal model are developed by the authors, which predict:

* A positive relationship between divisibility and number of contributions
* An ambiguous relationship between divisibility and the number of complete contributions (a complete contribution refers to members completing all the tasks that encompass a full contribution)
* An ambiguous relationship between divisibility and value of the contribution

## Evaluation of the Difference in Difference design.

A difference-in-difference model is applied by comparing contributions to the project “Cyclone Center” on Zooniverse, which implemented the aforementioned format change, to a control project “Galaxy Zoo” which was not affected.

For a difference-in-difference design to make sense, treatment has to affect a certain amount of individuals within a population at a specific point in time, while it doesn’t affect another portion of individuals within the population. Under certain assumption, this setting can be used to tease out the causal effect of a treatment by comparing, as the name suggests, the difference in differences between the two groups before and after the treatment period.

Such a case presented itself in this study, as within the same community, a project introduced a change in task divisibility by doubling the amount of steps it requires to complete a task (the treatment) while this was not implemented in a control project on the same platform.

A few points that add to the adequacy of the setting is the fact that changes were not announced prior to the implementation, which meant that there is no possibility of anticipatory effects within the treated set of users. Furthermore, the number of steps necessary to be undertaken for a complete edit varied systematically after the treatment was introduced. Depending on the task, between 5-8 steps (instead of 3) had to be completed to make a complete edit. This allowed the researchers to investigate whether the diff-in-diff estimator varied according to the number of additional steps that the change introduced. Finally, factors like contributors being randomly assigned to tasks as well as the (intuitively valid) assumptions that tasks do not differ in terms of interest to the contributor add to the suitability of a diff-in-diff design for this natural experiment.

### Is the control group a plausible control group

What does lend credibility to the control group is that they are both on the same platform which deals with a tight focus on crowd science. Both projects demand the same basic skills of image recognition and categorization without any deep prior knowledge in the field by the contributors and both have a similar workflow design.

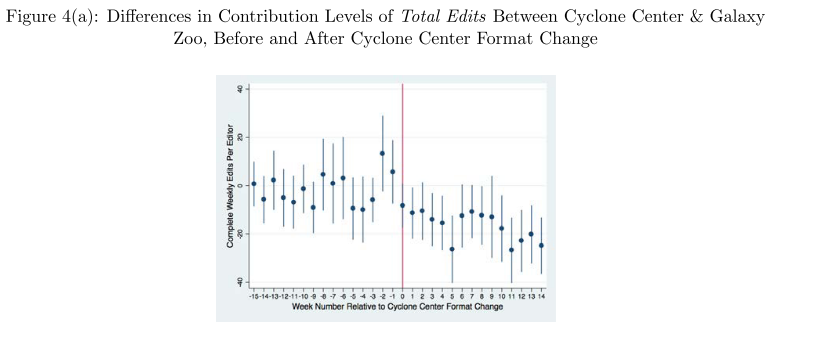
However, looking at the descriptive statistics, the control group is made up of approximately 4x more individual editors and 20x edits/contributions per week compared to the treatment group. There is grounds for suspicion that the smaller group size of the treatment group may be driving results, or that the tasks in the control group are more interesting (which is an issue as authors invoke intrinsic motivation as a theory)[[1]](#footnote-1)

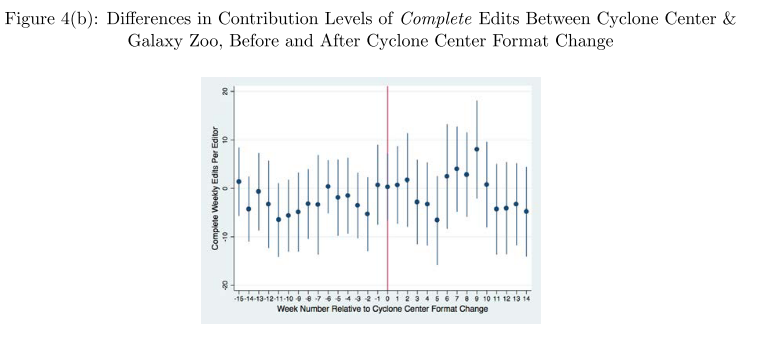
### Parallel trends evaluation

Observations are aggregated to a weekly basis, on which the parallel trends analysis is done with the following formula:



Significant coefficients on the interaction term in the formula above indicate that there are differences in the contribution levels between the treated community and the non-treated community in a given week. For the parallel trends assumption to hold there should be no significant coefficient on the interaction terms before the time of treatment. The figures below illustrates that the parallel trends assumptions holds in the weeks prior to the treatment introduction, as all of the significance intervals cross 0 on the y axis.





### Can the results be interpreted as causal:

The specification of treatment and control group has been made sufficiently clear. Without going into too much detail on how tasks on the platform are similar among both groups, there seems to be very little possibility for the two groups to differ substantially in their contribution behavior. Furthermore, the authors provide convincing evidence for a causal effect of the treatment by both demonstrating that the parallel trends assumption does hold and by adding an additional control group to provide robustness checks. Furthermore, concerns about other effects (that seem to be suggested by editors as I don‘t see how this should be time variant) like day-of-the-week confounders are addressed. Finally, the results seem intuitively consistent, as an increase in work load for the same pay off (completing a task) should reduce the amount of contributions that are made, which is reflected by both the formal model as well as the empirical results

The only minor improvements I would make are based on a prior lecture as well as my own research. First, I believe the plausibility of the control group could be improved by applying some form of matching procedure. I am not aware of variables that the authors may have access to on the level of individual contributors, but I believe it would add to the explanatory power if they could make some improvements to the similarity of the treatment and control group, even if the parallel trends assumption holds. Alternatively, adding in controls in the main specification that are time variant would make sense. Given the experience and competence of the authors, I doubt this is due to a knowledge gap and rather due to a lack of data on the individual level. In addition, I don‘t quite understand why they analyze count data using an OLS model, rather than a Poisson or Negative Binomial model. They don‘t give much information on the distribution of the standard errors but, at least according to my understanding, count data doesn‘t tend to be normally distributed, hence making the standard errors less efficient when using a plain vanilla OLS.

**Link to paper:** [**https://www.nber.org/papers/w23955**](https://www.nber.org/papers/w23955)

1. I noticed that in an updated version of the paper from 2019, a further control community was added as a robustness check as apparently concerns were raised that the control group project was in its fourth iteration, raising concerns about the a priori motivation of contributors to take part in a project that is in its fourth round. Furthermore, the size of the control group was reduced to more closely match the contribution behavior of the treatment community, however there was no information provided as to how this was accomplished. [↑](#footnote-ref-1)