Maryiam Haroon and Azam Chaudhry (2014). Where Do New Firms Locate? The Effects of Agglomeration on the Formation and Scale of Operations of New Firms in Punjab. Economics Discussion Papers, No 2014-21, Kiel Institute for the World Economy.
http://www.economics-ejournal.org/economics/discussionpapers/2014-21

Responses to Referee Report 2:

Comment 1: From the first sentence of the introduction, the reader is brought to think about the benefits of the formation of new firms. There is a difference between understanding whether new firms do appear, and why, and understanding where do new firms appear, conditionally on being created. Paper addresses the second question and should stick to this presentation.

The paper analyzes the second question about where new firms choose to locate, or more specifically whether they chose to locate in agglomerated districts or not. This can be made clearer in the paper.

Comment 2: The second sentence is also very misleading to me. The "well established socio-economic benefits" are not defined (until the next section), and come as if we had been discussing the issue for a long time. This relates to the direct effects of the creation of new firms on individuals of the neighborhood

The term "well established factors" has been mentioned in order to put forward the idea that there are employment opportunities created when new firms locate in a specific region. The purpose of the terminology was to give a general statement about the benefits of new firms though it can be omitted.

Comment 3: Then, the theoretical elements come into the discussion, which is nice. However the authors cite Marshall and do not cite Duranton and Puga, which have written very nice and relatively recent literature review on the theoretical explanations of agglomeration economies.

We can incorporate the Duranton and Puga contribution in the revised version of the paper since it provides a good overview of theoretical literature and analyzes different types of micro foundations of agglomeration economies such as sharing, matching and learning mechanisms.

Comment 4: Then, the Jacob-type of agglomeration economies come into the discussion, which are a type of agglomeration economies and could be brought forth somewhere else than in the introduction. The broad picture is missing.

We can also incorporate the sharing, matching and learning mechanisms discussed by Duranton and Puga in the introduction and move the Jacobs reference to later part of in the revised version.

Comment 5: Other empirical papers are cited, however we still do not know what the current paper does.

This paper investigates the location decision of firms within an industry with respect to the two agglomeration forces. Specifically it tests two hypotheses: the first one is that while making their location choice decision do new firms in an industry chose to locate in an area where there is presence of similar activity (which is referred to as localization)? Second, do new firms chose to locate in an area where there is presence of diverse activity (urbanization)? This can be made clearer in the introduction.

Comment 6: The paper addresses the question of agglomeration externalities, with a focus on learning externalities. Sometimes the learning externalities focus disappears in the paper.

The model in the paper develops the model through learning externalities and the presence of workers in an area while the empirical estimation focuses principally on presence of workers. Due to data limitations the empirical estimation is not directly testing for learning externalities so we avoid giving explicit explanations of the results through this channel. However, we can modify the results section to include a discussion on learning externalities.

Comment 7: There are several issues about the cited (and not cited) papers in the manuscript. First, in my view there is a lack of references to the existing literature in agglomeration, trade and location economics. Also, the cited references are sometimes misplaced. For example, Otsuka (2008) comes in the first page and appears as a leading paper, while there could be other papers placed before, linked to the current topic. Sorenson (2000) is a sociology paper. It should be said, both because the reference comes from outside the IO and trade literature, and because it is interesting to say that people in sociology study these types of questions. However the citation should maybe not figure as a main reference in the manuscript. Papers on location choices in the trade literature are for example *Crozet M., T. Mayer, J.--- L. Mucchielli 2004. "How Do Firms Agglomerate? A*

Study of FDI in France" Regional Science and Urban Economics Vol. 34 (1), January: 27---54.). **As a reader, I would like to see citations of the frontier of this type of empirical research, and also, more specifically, this type of research focused on learning externalities. The authors should explain the contribution of the location choice literature, and also the learning externalities literature. As a general comment I think the authors should take a step back from the exact question addressed in their paper, so as to be able to present their paper as part of the agglomeration and trade literature. Again, the current citations are not wrong, however they are to focused on "new firm formation". It would indeed by an interesting idea to explain to the reader the difference between new firm formation and location choices of firms. Finally some references do not figure in the Ref. section. Duranton and Puga for example. Some references do figure in this section and should maybe not. Where does the paper on contracting and efficiency in the surgical sector come into the discussion?**

The literature mentioned by the referee is important and relevant to the topic and we can include them and the other suggestions by the referee in the revised version of the paper.

Comment 8: The theoretical model by Soubeyran and Thisse is really interesting to read. However, the reader does not fully understand the necessity to have section 3 explain large parts of the model, when the empirical specification is based on equation 10. Instead of erasing the theoretical part, I would like to see the model better explained, i.e. to have the entire section 3 oriented towards obtaining a nice empirical specification.

Soubeyran and Thisse (2008) state that a higher number of workers and higher knowledge spillovers are the two factors that attract new firms to locate in a specific area. The estimated equations are connected to the number of workers because we are not able to incorporate the knowledge of workers or the technological innovations occurring due to data limitations. We subdivide the number of workers as in Soubeyran and Thisse (2008) into two components: the total number of workers in a district and the number of workers in a particular industry within a particular district. The existing model can be adjusted by adding the superscript on the number of workers (L_d) in equation 10. The superscript will vary in two ways: it will vary across the number of workers in a region and number of workers in an industry within a district. We moved from equation 10 to 11 and 12, since equation 10 states that higher number of workers will lead more new firms to enter in an area, which indicates that more employment will lead more new firms to enter. So, the number of workers (L_d) in equation 10 has been decomposed into two components in equation 11 and 12 and the two components are the number of in a particular industry in a district (which is referred to as localization) and the total employment in a district (which is referred to as urbanization).

Comment 9: The authors use data on Pakistanese firms. Some questions arise. Why only the years 2006 and 2010? The first paragraph in section 4 is not that clear. Is the database exhaustive? The sentence "We have used the DOI 2010 to measure the arrival of firms in 2008" is not clear. If the reverse causality issue is the main explanation, then the arguments should be made the other way around.

The data set used for the analysis is the Punjab Directory of Industries data which is available only for two time periods, 2006 and 2010. The dataset includes the information about all the firms in Punjab, Pakistan. Since, there can be reverse causality between arrival and agglomeration forces, we used lagged values of the independent variables (agglomeration forces). In other words, we measured agglomeration forces (which are the independent variables) in the time period 2006 and arrival (which is the dependent variable) is measured in the time period 2008 in order to account for causality. We can use the 2010 data set to find the new entrants in 2008 since the dataset includes information about the year of establishment of each firm so those firms whose year of establishment is 2008 are classified as arrivals in 2008.

Comment 10: More information should be given on the dataset. Are we sure that these are firm creations? Can't these be firm movements? Before showing us data on the potential urbanization and localization issues, we would like to see comparisons across years, for example. Maybe references to other countries. Again, here the reader should understand why the "firm creation" issue is specific and different from an increase in employment, for example, or a multinational firm's location choice. I could be interested in asking whether the firms which increase their employment in the considered period, are more specifically located in the agglomerated areas. Is this question different from the question addressed by the authors?

The dataset includes more than 18000 firms and there are 34 districts according to the classification. The summary statistics of the variables are included in table 1 and 2. Unfortunately, we do not have access to the data set for other countries. The dataset does not report whether a new firm is a movement or new establishment and that is a limitation of the data. Our paper does not test whether firms have increased their employment levels due to agglomeration but rather what is the impact of localization and urbanization on the scale of operation which is measured by total employment levels of new firms. Specifically, the paper tests how a firm's scale of operation (measured by total employment) is related to the presence of similar activity (localization) and also how it is affected by the presence of small scale firms, medium size firms and large scale firms (disaggregated localization).

Comment 11: The two specifications are very interesting questions. However, they need to be placed in the existing literature. For instance, I don't see where the demand effect is addressed, nor the competition effect.

The literature suggests that the benefits of agglomeration are greater than the costs associated with locating in agglomerated areas. The paper does check for the significance of non-linearities in order to test for the competition effect which can be made clearer in the paper. The paper also includes other characteristics of the district such as the average age of the population, the percentage of male population and the average income and these factors can account for the demand effect. Again this can be made clearer in the paper.

Comment 12: Why not analyze the location choice at the firm level? The dataset is at the firm-level.

It may not be appropriate to estimate the model at the firm level because the dependent variable may be skewed since we will only have the data for firms which entered in a specific year and we do not know about the firms that did not enter. The analysis at firm level will be more appropriate if we look at the exit of firms since we can examine the firms that exit and the ones that entered which can be done as an extension to this paper.

Comment 13: Why do we have only one year? Using the panel could allow to control for all the regional characteristics.

The analysis is for single year because the district level socio-economic data was incomplete for the other years and we are planning to make to a panel for future papers using different time periods.

Comment 14: The authors should explain which is the variability which explains the estimation. The explained variable is the total number of new firms in industry i and district d in a single year. Is the only variability here, for a given district, among the different industries, and for a given industry, among the different districts? The authors should be clearer about the interpretations they want to push forward.

The dependent variable is the total number of firms in industry i and district d. We are looking at the second case which is for a given industry or in other words we are testing for variability among districts.

Comment 15: The explanation of the two estimated equation begin by presenting the error terms. This is not adequate.

The discussion of the error terms can be moved.

Comment 16: The expression "socio –economic factors" is repeatedly used. This should be avoided, a well as for other often repeated "ready to go" expressions.

The revised version will take into account the suggestion.

Comment 17: The results section is nicely explained. However it could be a little longer, with an analysis which could take a step back. Comparing with other existing studies is interesting, but I don't see papers in the trade literature here. Also, results should be discussed differently. Selection effect is very important. Say more on this. Say more on oligopolistic type of behavior also.

We can incorporate a discussion on the learning mechanism in the results section in order to relate it to learning externalities.

Comment 18: How does the model by Soubeyran and Thisse intervene in shaping these results? Is it necessary to have the specifications taken out of this model specifically? This is really important to add.

We have explained above how we can modify the Soubeyran and Thisse model in order to explain why we estimated the empirical model.

Comment 19: There seem to be a problem with section 6.1. Why is the robustness section so short? Also, checking the same estimations without controls does not really represent robustness checks.

We kept the robustness section brief as we already estimated the model using independent variables measured in 2004 as well and the results of that are presented in the main result section. The robustness estimations drop all the variables which were used to control for district characteristics such as socio-economic characteristics of a district but control for district level effects by including district dummies. This can be made clearer in the discussion of results.