Review of: "Digitalization of research: do ICT improve scientific production in developing countries?"

Concepta McManus[1]

1 Universidade de Brasília

Potential competing interests: No potential competing interests to declare.

The paper looks at Information and Communications Technology and how this can affect scientific production in developing countries, and as such, it is an interesting topic. Nevertheless, the paper needs significant changes. First of all, it does not explain what ICT means or how it was computed.

References are, in general, very old. They should be limited to 5 or, at most, 10 years. The introduction needs to bring the reader to a clear objective of the study.

Material and Methods needs a lot of work. You say, "The variables ICTit are proxies of ICT diffusion, including the Internet and mobile phones." This says little. Later you say, "The main explanatory variable is the use of the internet. We capture it by the proportion of individuals using the internet." Is this your internet variable? What about mobile phones, which were mentioned earlier? All abbreviations should be defined clearly, which is not the case here. What developing countries - Sudan is spelt wrong, CAD is not specified, there are only 69?

How were the variables measured? First of all, all variables need to be carefully defined - what are the variables in the second equation? Corr, tertiary_educ, ICRG, etc. (tertiary is spelt wrong), what were their time frames, etc. What is mining rent? What time frames are used for each variable? If they are not the time frames, it should be remembered that determinant variables are only relevant if they come before the dependent variable. What collinearity exists between the independent variables? If there is high collinearity, how can conclusions be made? Why take the log of the GDP? - this all needs to be explained. The authors say they use data from different reports and databases. They need to explain what data specifically and how it was used, as well as the year. For example, they say they use the World Happiness Index, and this only began in 2011, but they say their data is from 2000-2016.

Results: What are 1, 2, 3, and 4 in Table 1? Baseline results of what? In Table 4, there appears a variable called Gov_stability, as well as internal conflict, lgdppc, among others. What are these? It should be noted that abbreviations change between tables, making the analysis very confusing for the reader. This also complicates the discussion section. All Tables should be able to stand alone without textual explanations. What are the values in the Tables? This is not clear. Definitions of variables, etc., are introduced in the results section. It also seems that the paper is not completed as there are questions asked and not answered in the text ¬ How did you handle the issue of missing data? As seen in Table 1, they are missing information. The number of codes is less than 70, and you failed to present the number of instruments¬ This also complicates our ability to give suggestions for the paper.   Having both *** and the level of

significance in parentheses overcomplicates the tables. What are brevets? Table 4. The definition of levels of significance is * *P<0.05,* **P < 0.01, and *** P<0.001, and not as used here.

The discussion suffers from the lack of earlier clear definitions of variables measured. For example ¬ the quality of public education policies positively and significantly explains the number of patents filed and the number of scientific articles published. ¬ What public education policy? There was a mention of the number of women in tertiary education, but what POLICIES?

Another comment ¬ Second, national health expenditures negatively explain the productivity of researchers in developing countries. This is because, due to the low life expectancy prevailing in developing countries, these countries still allocate too few resources to health financing.¬ is confusing. An increase in health expenditure decreases productivity. This is because people die early since there is too little financing. This makes no sense to me.

There are regional analyses - not presented in the materials and methods - which seem interesting but, once again, lack clear definition, and results are not clearly presented by region - what definition of region and how they were analyzed.

The Supplementary Tables are confusing - how are there 1190 observations (obviously NOT the number of papers) for 69 countries? What are these? - Also, each variable has a different number of observations - how was this accounted for in the model? Was imputation used? Correlation cannot be 10 - it must be 1.

Overall, my suggestion is a thorough review of the paper, updating references, defining variables and timelines, as well as providing clear statistical methodology to support the results before a clear discussion is made. The title, objective, and conclusions should be well aligned.