Referee report on "Biased Sanctions? Methodological Change in Economic Sanction Reconsidered and its Implications" by Peter A.G. van Bergeijk and Muhammad Shahadat Hossain Siddiquee

This paper proposes a critical evaluation of the change in coding of a database on economic sanctions developed by the Peterson Institute (PI). The PI has issued three editions with the core difference between the different editions being the coding of the data. The present paper demonstrates with the support of case studies, and with an econometric analysis, that the changes across the different editions are far from being innocuous since the degree to which the sanctions are expected to achieve the desired goal has changed for coding reasons alone. The authors of the present paper criticise the fact that the changes in the definition of sanction, and therefore in coding, have not been transparent and justified by the PI.

The points raised by the authors are broadly valid, but I am not sure that this paper is optimally framed. It is not surprising that the modification of a definition and of coding will generate changes in the statistical analyses conducted with these databases. As such, too much space is being devoted in this paper in showing this specific (and obvious) point. At the end of the day, it does not matter whether sanctions become more or less effective after the change of coding. What really matters - and this point is certainly underdeveloped in the manuscript - is whether the new definitions of sanctions are more meaningful than the previous ones. If the new definitions are more accurate, then there is not much scope for criticism since (as the authors themselves acknowledge) that's how science evolves. On the other hand, if the new definitions are biased, then the authors should focus their analysis on the deficiencies of the new definitions.

Comments

- p.6 (please number the pages!), last paragraph: it is stated that the change in definition increases the success score for case 48-5 from 4 to 6, and the authors conclude that this is a 50% increase. While this is a fact, it is not very informative. The points are on a 16 points scale. Whether the increase is from 1 to 3 (300% increase) 4 to 6 (50%), or 10 to 12 (20%), it is not the percentage changes that matter, but the absolute points change. The same comment applies to subsequent paragraphs.

- The case studies are not well explained. Unless one is really familiar with the details of the cases being analysed, it is not possible to fully comprehend this section. If the authors believe the case studies should be maintained in the paper, I would therefore suggest them to be much clearer, and to provide a thorough description of the analysed sanctions.

  In section 3.1 the authors detail the changes implemented in the definition of the measurement of success from the 1st to the 2nd and 3rd editions. The changes seem very reasonable to me, to the extent that the new coding allows a more precise decomposition of different degrees of success. I therefore do not grasp the criticism of the authors here. In other words, and reiterating what I said earlier, the definition change should not be interpreted in light of the change in the distribution of point (as the authors do) but instead in terms of the meaningfulness of the new definition as compared to the previous one. Does the change in definition constitute an improvement (as it seems to me) or not? The focus should be on this alone in my view.

  In the subsection *involved countries* the same comment to above applies. The coding of involved countries has changed from one version to another. But does the latter edition feature a more accurate classification of involved countries, or not? If the answer is negative, then the authors should present these cases in detail and make a point as to why this is not meaningful. Otherwise, I see no scope for criticism since the newest edition would be improved compared to earlier versions

In Table 5, columns 7 and 8, the authors use the full sample for the 3rd edition, but this is not meaningful. In the 3rd edition, the PI has actually doubled the sample size by distinguishing the sanction emitter from the sanctioned country. Since this is being coded, this informations should be used in the regression, and to avoid comparing apples to oranges, the double counting should be avoided by keeping the observation that is the most meaningful given the identification strategy.