

Dear Referee,

First, I would like to express my great appreciation to you for your response and very constructive comments. The comments and objections are quite useful and I believe they will greatly improve and strengthen the paper. I am currently revising the paper in light of your comments and will upload the revised version after a period of time. Please find below my quick reflections on the major issues.

As the referee rightly pointed out, the model specification and empirical results were presented in a rather terse style. Specifically, I did not report details of important test statistics, dummy specification (for some of the models) and computation of impulse response functions. In addition, the paper fails to point out upfront the caveats associated with results based on a very small sample such as ours. In addition, the Referee has raised important methodological issues, most of which were not explicitly discussed in the paper. Below are my point-by-point responses to the major issues raised by the Referee. Note that the minor comments are not discussed here but they will be fully taken into account when revising the paper. Please let me know if any of the points discussed here happen to be unclear or if you have any further questions.

Response to major comments

(i) Methodological choices/transparency

- The revised version will have a table presenting detailed specification test results (including values for excess kurtosis and skewness, autocorrelation, cointegration rank) for all three models. Concerning the test for normality, in all three models, the assumption of multivariate normality is only borderline accepted (details will be forthcoming shortly). However, a look at the univariate test statistic indicates that normality was not rejected for all equations, albeit with only a small p -value for some of the variables, which is due to excess kurtosis in the equations. However, all individual equations have a skewness close to zero. We have gone to great lengths to ensure a model setup where multivariate normality is accepted with a higher p -value by, among others, estimating a partial model conditioning on weakly exogenous variables and changing the sample period; however, all these avenues lead to similar conclusions. As noted in the paper, no serious deviations from the assumption of normality and residual independence has been detected; however, there are some indications of excess tails (long tails) and moderate ARCH effects. However, because cointegrated VAR results should be reasonably robust to excess kurtosis and ARCH effects (Gonzalo, 1994; Rahbek *et al.*, 2002; Juselius, 2006), we regard the present model specifications to be acceptable.
- As rightly mentioned by the Referee, a correct model specification (including values for normality tests, trends and dummies) is critical as it affects the (asymptotic) distribution of the rank test. It is, however, important to note that, given the small sample at our disposal, the trace test is likely to yield unreliable results due to both *size* and *power* problems. Although the small sample Bartlett correction provides a correct size, it does not necessarily circumvent the power problem (Johansen, 2002; Juselius, 2006). So, anyway, we should not focus much on the trace test for an analysis based on a small number of observations. From my own experience with VAR models, even a sample with 50-60 observations is too small for the trace test to be a reliable indicator of the cointegration rank. This is why it is crucial to complement the formal testing (using the trace test) with other indicators of rank (see discussion in Juselius (2006: Chapter 8)). Accordingly, we have made the choice of the cointegration rank based on a range of (additional) statistical criteria, such as the largest unrestricted root of the characteristic polynomial, the t -ratios (significance) of

the adjustment coefficients (α); and the graphs of the r^{th} cointegration relation. However, I check whether the robustness of the results based on the *first-best* choice of rank are sufficiently robust to changing the rank to the *second-best* choice (these results will be reported in the revised version). The robustness check indicates that the key results stand up well to changing the rank value.

- The dummy variable for 1989 in Model 1 is clearly a mistake as the sample period starts in 1990. I thank the Referee for bringing this to my attention. In addition, as the Referee observed, I did not explicitly discuss dummy specification and, in particular, the types of dummy the extraordinary observations were accounted with. This will be addressed in the revised version.
- The Referee “failed to replicate the results of Model 1.” But, as already mentioned above, the details of model specification were not sufficiently discussed in the paper. This is probably the reason why the Referee could not replicate the results for Model 1. In other words, differences in results are likely due to differences in deterministic specifications.
- Yes, the identifying assumptions underlying the computation of the impulse response functions in Section 6 were not explicitly discussed in the paper and will be addressed in the revised version.

(ii) Adequacy of the data

- The sample size is admittedly very short in view of the large number of parameters. I completely agree with the Referee’s comment that the paper should discuss upfront the caveats associated with this and that the empirical results need to be interpreted keeping these in mind. However, we used data spanning the period 1990-2014 for Model 1 and 1980-2014 for Model 2 for the following reasons. *First*, Tanzania’s national accounts data for the period before 1980 are not reliable. *Second*, as stated in the paper, Model 1 uses data for the period since 1990 because Tanzania’s exports to China were negligible in the 1980s. China’s strong economic ties with Tanzania are a relatively recent phenomenon and thus analysis based on a sample period that also covers the 1980s and 1970s would be less valid and less meaningful. Given that the Cointegrated VAR model exploits the variation in the data over time, it would also be econometrically implausible to include the period before the early 1990s since Tanzania’s exports to China during this period accounted for a negligible amount and featured barely any variation. For these reasons, Model 1 focuses on the period since 1990. An alternative would be to use higher frequency data and focus on the period since 2000. However, quarterly or monthly data on the macro-variables in the analysis are nonexistent. Therefore, we tried to deal with the problem by including only 4 variables and modelling some of the variables as weakly (long-run) exogenous, which enabled us to preserve some degrees of freedom. However, for SSA economies, analysis based on a short sample is not uncommon in the literature (see, for instance, Juselius *et al.* (2014a,b) and Gebregziabher (2014, 2015)). Having said these, I want to emphasize that the claims based on the results in the paper will be made less forceful and will be complemented with a stronger narrative and theoretical arguments.

(iii) Plausibility of the results

- I would say the results are not implausible in light of the findings of other studies. Our results are consistent with, for instance, the finding in Drummond and Liu (2013), who show that a 1 percentage increase in China’s domestic investment growth is correlated with an average 0.6 percentage point increase in Sub-Saharan Africa (SSA) countries’ export growth. In particular, focusing on the top five resource-rich SSA countries (Angola, South Africa, the Republic of Congo, Equatorial Guinea, and the Democratic Republic of Congo), they show that a 1 percentage point slower investment growth in China is associated with 0.8 percentage points lower export growth: the more resource rich a country is, the larger the impact on export growth. Note that the majority

of SSA countries in Drummond and Liu (2013) exported only less than 15 percent of their exports during 2001-2011 to China. It should also be kept in mind that, for many of these countries, the shares of exports to China were much smaller in the 1990s than in the 2000s. Given that the above-mentioned study also looks at the impact of China's domestic investment on the exports of SSA countries and focuses on more or less similar period, I would think our results are not implausible.

- I acknowledge that the analysis does not include all relevant variables and some important variables are omitted. In particular, as indicated by the Referee, China's domestic investment may be capturing "features of the world economy that are not captured by World GDP." It is worth noting, however, that although our analysis omits important variables, this does not, in general, invalidate the long-run estimates. The reason is that cointegration property is invariant to extensions of the information set, i.e. a long-run relation detected within a given set of variables will also be found in an enlarged variable set (Johansen, 2000; Juselius, 2006). Put another way, inference on the long-run structure still remains valid in a system with smaller set of variables.
- Having said these, I will try to conduct more sensitivity analysis to check the robustness of our key results, along the lines suggested by the Referee.

(iii) Conceptual issues

- **Parameter constancy:** With only 25-35 observations, no serious check for parameter constancy can be carried out. Therefore, some of the estimated coefficients in our models may represent average historical effects over the sample period. However, parameter instability is likely to be less pronounced given that we have appropriately accounted for the most significant events that could have engendered parameter non-constancy. It should be noted that analysis based on a very short sample (and grounded on the assumption of parameter constancy) is not uncommon for SSA economies (see, for instance, Juselius *et al.* (2014a,b) and Gebregziabher (2014, 2015)).

References

- Gebregziabher, F. H. (2014). The long-run macroeconomic effects of aid and disaggregated aid in Ethiopia. *Journal of International Development*, 26(4): 520-540.
- Gebregziabher, F. H. (2015). Adjustment and Long-Run Economic Performance in 18 African Countries. *Journal of International Development*, 27(2): 170-196.
- Gonzalo, J. (1994). Five alternative methods of estimating long-run equilibrium relationships. *Journal of Econometrics*, 60: 203-233.
- Johansen, S. (2000). Modelling of cointegration in the vector autoregressive model. *Economic Modelling*, 17: 359-373.
- Johansen, S. (2002). A small sample correction of the test for cointegrating rank in the vector autoregressive model. *Econometrica*, 70: 1929-1961.
- Juselius, K. (2006). *The cointegrated VAR model: Methodology and applications*. Oxford: Oxford University Press.
- Juselius, K., Reshid, A., and Tarp, F. (2014b). The real exchange rate, foreign aid and macroeconomic transmission mechanisms in Tanzania and Ghana. Discussion Papers, University of Copenhagen.

Juselius, K., Møller N., and Tarp, F. (2014). The long-run impact of foreign aid in 36 African countries: Insights from multivariate time series analysis. *Oxford Bulletin of Economics and Statistics*, 76(2): 153-184.