CREDIT Research Paper

No. 15/06

Foreign Aid, Poor Data, and the Fragility of Macroeconomic Inference

by

Lionel Roger

Abstract

The link between foreign aid and economic growth remains a controversial issue in the literature, and a large share of the disagreement could be explained by differences in the data employed. Using GDP data from three different versions of the Penn World Table and the World Development Indicators, I investigate the robustness of Juselius, Møller and Tarp (2014)'s (JMT) conclusions about long-run aid effectiveness. The analysis is carried out in two stages. First, I apply the same models as developed by JMT to the new datasets. Second, I re-specify the Cointegrated VAR models using the same criteria as JMT, but limit the analysis to the four most and least consistent countries respectively. The first exercise shows that results change in a significant manner in approximately 10 of the 36 countries examined. The second exercise shows that the if the models are re-specified for each country and dataset individually as a function of the data, this leads to more qualitative changes in the conclusions.

JEL Classification: C32, F35, O11, O47

Keywords: Time-Series Models, Foreign Aid, Economic Growth

Centre for Research in Economic Development and International Trade, University of Nottingham

CREDIT Research Paper

No. 15/06

Foreign Aid, Poor Data, and the Fragility of Macroeconomic Inference

by

Lionel Roger

Outline

The Author

Lionel Roger is a PhD candidate in Economics at the University of Nottingham. Contact email: lexljrog@nottingham.ac.uk (replication files available there upon request).

Acknowledgements

I am thankful to Oliver Morrissey and Markus Eberhardt for their helpful comments, and to Niels Framroze Møller for providing me some of the data and programs underlying Juselius, Møller and Tarp (2014). This work was supported by the Economic and Social Research Council [award number 1511835].

1. Introduction

The question whether foreign aid has a positive effect on recipient economies remains a controversial issue in the empirical literature. While many studies conclude that, on balance, aid has a beneficial impact on economic performance (e.g. Hansen and Tarp, 2001; Brückner, 2013; Temple and Van de Sijpe, 2014), others come to the conclusion that it is harmful (e.g. Rajan and Subramanian, 2008, 2011) or that it has no discernible effect at all (e.g. Easterly, 2003). In their recent contribution to this debate, Juselius, Møller and Tarp (2014a), henceforth JMT, address many methodological issues of the previous literature by applying the cointegrated VAR (CVAR) framework to a sample of 36 sub-Saharan African countries. Their study provides strong evidence in favour of the hypothesis of aid effectiveness, finding that aid has had a positive long-run effect on either Investment or GDP itself in 27 of the countries in their sample. From the outset, they rule out the possibility that differences in the sources of the data could play a major role in explaining the often contradictory results of the literature, since "most studies use data from the exact same publicly available databases" (JMT, p. 1). The differences would therefore necessarily go back to methodological differences between the studies. This view however, is not put to test in their study, which is based on one dataset, namely the Penn World Table v. 6.3.

Although the focus of this paper is the implications for statistical inference of using alternative measures of real GDP on the same country sample period and with the same estimation method, it relates to broader literatures. Testing robustness to alternative data is one way of addressing concerns about unreliable GDP estimates for developing countries, especially for Africa (Jerven, 2013c), that have recently attracted much attention from both policymakers and the media. A related strand of literature identifies substantial divergence between the most commonly used datasets, and shows how this can often have a fundamental impact on the macroeconomic inference that is based on them (e.g. Johnson et al., 2012; Ponomareva and Katayama, 2010). The analysis is also relevant to the literature on aid effectiveness, in particular concerns over replication and robustness. A number of studies have demonstrated the sensitivity of results to even minor changes in the underlying data (Easterly et al., 2004; Roodman, 2007), although typically these refer to changes in the sample and/or estimation method (both of which are held constant in this study).

Time series analysis approaches to the question of aid effectiveness, such as the one taken by JMT and a growing number of other country-specific studies have many merits, as they are powerful in accounting for the substantial heterogeneity between countries, and can exploit the full potential of the time series properties of the data, while being robust to its pitfalls. However, they typically put high requirements on the quality of the data, and the high level of technicality and the time consuming process of model specification can act as a barrier when it comes to performing robustness checks. It is therefore understandably not common practice in this branch of literature to assess the robustness of conclusions to alternative datasets, and to the best of my knowledge, no study has been dedicated to determine the sensitivity of this type of investigation to changes in the underlying data, or to alternative datasets providing different measures for the same core variables.

I fill this gap, exploiting the fact that JMT offer an excellent framework by providing economically and statistically sound time series models for a set of 36 sub-Saharan countries, and a well grounded procedure of model specification that can serve as a blueprint for potentially due re-specifications as the data changes. I will apply this framework to three widely used datasets, namely the Penn World Tables versions 7.1 and 8.0, and the World Bank's World Development Indicators.

I proceed in two stages: First, I apply the exact models as specified by JMT to the new datasets, ignoring potential implications of differences in the data for the modelling process. Second, while maintaining the same econometric framework as JMT, I re-specify the variable components of the models as a function of each dataset individually, for selected countries. This is consistent with the approach taken by JMT, and more generally, in line with the philosophy of the CVAR approach of "allowing the data to speak freely" (Hoover et al., 2008).

I show that for approximately one third of the countries in the sample, the results change qualitatively, while two thirds remain stable. Endogenising the model specifications, that is, individually constructing them as a function of the respective data, generally leads to a greater divergence in the results, but significantly less so for countries where the results remain stable in the first exercise.

The remainder of the paper proceeds as follows: Section 2 provides a brief overview of the literature, focussing on the strand of the aid effectiveness literature that employs methods of time series analysis, and of the literature on the quality and consistency of macroeconomic data. Section 3 introduces the datasets and discusses the extent and nature of their divergence. Section 4 briefly introduces the CVAR methodology employed, and in section 5 I specify my replication procedures and present the results of both exercises. Section 6 concludes.

2. Literature review

The study at hand is situated at the intersection of two strands of literature, namely that on aid effectiveness, in particular that using time series methods, and that on the quality of macroeconomic data. This section will provide a brief overview of each of these, highlighting the aspects that are of particular relevance for the subsequent discussion.

2.1. Aid effectiveness

There is a vast literature on aid effectiveness at the macroeconomic level. It has been following a path similar to the broader empirical growth literature, with early studies typically relying on cross-section data (e.g., Boone (1996)), and later panel data to exploit the temporal variation in order to estimate the effect of foreign aid on economic development (e.g., Dollar and Easterly (1999); Burnside and Dollar (2000); Dalgaard et al. (2004)). Overviews of this early but still influential literature are provided by Hansen and Tarp (2001) and Morrissey (2001).

One common criticism about these approaches, however, is that their methodologies inherently imply relative strong assumptions about the homogeneity of the countries included in the sample (see Temple, 2010, pp. 4503). Given the great institutional and economic diversity among low-income countries, it is likely that the channels that link foreign aid and long-run growth, are substantially different in one country from another, which is masked by the estimation of *common* parameters for all countries[1]. Second, even the panel approaches do typically only exploit temporal variation in a very rudimentary manner, averaging over periods of typically 5 years, and widely ignoring the time series properties of the data.

In order to address these issues, a time series approach to the measurement of foreign aid effectiveness has become popular in more recent years. First, by focussing on country-specific studies, with individual statistical models, this literature addresses the issue of heterogeneity in a radical manner. Second, by exploiting and adequately addressing the whole extent of the time series dimension of the data, this literature is able to make credible claims about the long-run and short-run relationship between macroeconomic aggregates.

In an attempt to estimate the fiscal effects of foreign aid in Ghana, Osei et al. (2005)

[1]This issue has recently been addressed by Herzer and Morrissey (2013), who estimate country-specific parameters using heterogeneous panel cointegration techniques.

take a CVAR approach, exploiting inter-temporal variation in the fiscal variables and annual aid flows. They find that foreign aid generally had positive effects on the government's fiscal behaviour, reducing domestic borrowing and increasing tax effort.

Martins (2010) undertakes a similar exercise for Ethiopia, exploiting quarterly data over 15 years. In line with Osei et al. (2005), he finds that aid has no adverse effects on tax effort, and that it is mainly used as an alternative to domestic borrowing.

In a case study of Kenya, Morrissey et al. (2007) address the question of the fiscal impact of aid, as well as its impact on economic growth in general, in turn relying on a version of the CVAR. By disaggregating aid flows, they find that while loans tended to be used in order to fill unexpected fiscal gaps, undermining fiscal discipline, grants lead to increased public spending without reducing tax effort or crowding out private investment.

While all of these studies have the merit of circumventing many issues of heterogeneity by exploiting information contained in the temporal interaction of the variables of interest, country-specific studies come with the limitation of being very specific, and only have very limited generalisability. Juselius et al. (2014a), the benchmark study in the paper at hand, address this with a study comprising individual CVAR analyses for as many as 36 sub-Saharan African states, thus allowing for a substantial amount of heterogeneity while basing its final inference on a relatively large sample, comprising the majority of sub-Saharan African countries. They find that aid can be found to have significant positive effects on GDP in 17 countries, and detrimental effects in only 6 cases. When looking at investment, generally seen as the main driver of long-run economic growth, a positive effect can be discerned in 24 cases, a negative one in only 5. A positive effect on either of the two variables, by JMT's definition the indicator of aid effectiveness, can be discerned in 27 out of 36 countries. Their results therefore offer strong support to the hypothesis of aid effectiveness.

The study also sparked some follow-up studies, investigating in a more in-depth manner countries that stood out in the original study. Gebregziabher (2014) analyses the case of Ethiopia, using aid flows disaggregated by donor type (multilateral versus bilateral) and disbursement type (grant versus loan). He finds that, overall, aid has a positive effect on imports, GDP and investment, but reduces government consumption. However, these effects diverge greatly depending on the type of aid that is being looked at, highlighting the importance of disaggregation, that is, scratching beyond the surface, the main merit of country-specific case studies.

Singling out Ghana and Tanzania, Juselius, Reshid and Tarp (2014b) undertake a comparable exercise for two cases that stood out in the original JMT study, in the sense that aid appeared to have had ambiguous effects on their economic development. Their focus is on monetary and external factors, which they investigate by including nominal inflation and real exchange rates respectively in the equation, allowing for a much more differentiated analysis of the impact of foreign aid on these economies. While the overall conclusion about aid effectiveness becomes more nuanced, in that both countries benefited from foreign aid at least in some dimension, the results also indicate that both countries suffered from Dutch disease effects in the period under study, one of the main theoretical channels of aid harmfulness.

The approach taken by JMT offers a useful framework to assess the robustness of the inference drawn from this type of analysis with respect to the underlying macroeconomic data, as it provides carefully specified country-specific time series models for a large sample of countries, which can systematically be applied to data from various sources. Furthermore, it provides a sound framework for specifying such models, in case this is indicated by differences in the underlying data. The next section will provide a justification for the relevance of such an exercise, by outlining the results about both the divergence of such data, and the implication this can have on the inference of macroeconomic studies.

2.2. The quality of macroeconomic data

In a series of articles, Morten Jerven recently highlighted the deficiencies of macroeconomic data, especially in the case of African economies (Jerven, 2013b,a, 2011). The resulting book directed to a larger audience, "Poor Numbers" (Jerven, 2013c), has drawn considerable attention to the shortcomings of data collection. The main argument here is that, due to a lack of statistical capacity, political will, and changing academic currents, GDP measures quantifying the economic performance of African economies tend to be severely flawed, and often incomplete. While the study at hand does not touch on the quality of the original data, it focusses on the second-order implication that data providers, aiming to provide coherent and complete data for a large amount of countries and a long period of time, have to address the shortcomings of the original data. As the quality and the completeness of the original data decreases, the necessity for data supplying institutions to harmonise the measures and occasionally to account for periods of missing data increases, with methodological differences potentially creating discord between the datasets. Empirical studies relying on historical GDP estimates of low-income countries, the aid effectiveness literature included, almost exclusively base their results on these estimates.

The extent to which the different data sets tend to disagree is illustrated in a number of recent publications. Comparing the two most commonly used databases of internationally comparable GDP measures, the Penn World Table (in this case, version 7.1) and the World Bank's World Development Indicators, both based on essentially the same underlying National Accounts and price data, Ram and Ural (2013) show that the estimated GDP per capita in 2005 diverges by more than 25% in as many as 33 countries. The differences between the datasets range from -54 to +66%. In line with Jerven's line of argument, the largest relative differences (relative to GDP level) almost exclusively occur in low-income countries, and no developed country exhibits differences of more than 25%. It is worth noting that these differences do not follow an obvious pattern with respect to the datasets, in the sense that none of the sources systematically reports higher or lower incomes.

Significant divergence between different sources does not only occur between different data providers, but even across different vintages of the same series. Analysing the differences between four different versions of the Penn World Table, Ponomareva and Katayama (2010) find substantial divergence in the GDP growth rates. Using these four datasets, they replicate the influential contribution by Ramey and Ramey (1995), investigating the link between business cycle volatility and economic growth. The results vary strongly from one dataset to another and the main result (countries with higher volatility have lower growth) is not supported by some versions of the PWT.

A more systematic investigation of the differences between different versions of the PWT is carried out by Johnson et al. (2012), who compare the growth rates reported in versions 6.1 and 6.2. Their analysis shows that the divergence between the versions depends mainly on the size of the country, with smaller population numbers being associated with larger discrepancies in the data, and the distance to the benchmark year, i.e. there is more discrepancy, the further the benchmark year is in the past. They then perform the robustness check of applying the two datasets to 13 major empirical contributions to the growth literature. Besides the fact that the results do, overall, show a large amount of variability, their results suggest that this fragility is systematically linked to methodological properties of the studies. Crucially, studies relying on relatively long-term averages (over periods longer than 5 years) are relatively stable to changes in the underlying data, while dynamic analyses exploiting annual variation in the data yield the least robust results. This condition is exacerbated when the analysis refers to non-OECD countries, which, on average, tend to have lower quality data (or more divergent data between the datasets). This is of particular relevance for the study at hand, as JMT's analysis, and therefore mine, is both a dynamic one exploiting annual data, and focusses on some of poorest countries in the world, many of which are small, and typically have a comparatively low statistical capacity.

Comparing two more recent releases of the Penn World Table (versions 6.3 and 7.0), Breton (2012) finds even larger differences then those observed by Johnson et al. (2012). As the main driver of these differences, Breton identifies the (undocumented) methodological innovation of PWT 7.0 that discards all previous price benchmarks from 1970 to 1996, and entirely relies on prices measured in the 2005 round of the International Price Comparison Program (ICP).

It is therefore an established finding that there is substantial, often unsystematic and not always traceable divergence not only between different sources of macroeconomic data, but also between different vintages of the same source, or at least between the versions of the PWT. There are furthermore plenty of examples where influential empirical results turn out to be highly sensitive to these differences. Previous results suggest that a dynamic study exploiting annual variation in the data and looking at countries with typically low statistical capacity is susceptible to such fragility, all of which applies to JMT and therefore motivates my investigation.

3. Datasets

This study utilises data on GDP and expenditure shares (investment, household consumption and government consumption shares) from three versions of the Penn World Table (6.3, 7.1 and 8.0), as well as the World Bank's World Development Indicators (retrieved in March 2015), referred to as PWT6, PWT7, PWT8 and WDI in what follows. While all these measures are conceptually similar, the underlying methodologies and consequently the values that are reported by them – consistent with the findings of the earlier literature outlined in the previous section – sometimes differ quite substantially. All cover the sample periods used in JMT, but not necessarily for all countries, reducing the sample to 33 in PWT8 and to 13 in WDI. In this section, I will first briefly discuss the methodological differences and similarities between them, and then explore how this is reflected in the data.

3.1. Conceptual differences and similarities

Penn World Table 6.3 (Heston et al., 2009), the dataset underlying JMT's original study, is my benchmark dataset and serves the purpose of replication. The series JMT employ is real[2] GDP at constant 2005 prices, computed using a Laspeyres index (labelled RGDPL). The prices underlying the PPP adjustments in PWT6 are based on prices estimated in the ICP 1996.

Penn World Table 7.1 (Heston et al., 2012) provides the same variable, RGDPL. While this measure is conceptually equivalent in being real (PPP adjusted) GDP computed using a Laspeyres index, it differs in at least two ways: First, the Penn World Table 7 exclusively relies on the prices from the newer ICP round 2005. As noted by Breton (2012), it discards all older price data, inducing major differences in the reported growth rates. Second, the underlying concept of consumption is different from that in earlier (and later) versions of the PWT. Instead of differentiating between household consumption expenditure (HCE) and government expenditure, PWT uses the concept of *actual individual consumption* (AIC) and *collective government consumption* (CGC). The difference between the two lies in the treatment of goods and services that are consumed by individuals, but often paid for by the government, such as health care and education. AIC includes these expenditures, whereas HCE only includes such expenditures that are

[2]Note that *real* stands for *PPP adjusted* in the PWT.

actually being paid for by the individual. Even though AIC clearly has advantages when it comes to international comparability (by abstracting from the mode of provision and quantifying arguably equivalent consumption equally), it has been discarded in later versions of the PWT as the required data are not readily available for most countries, increasing guesswork.

The Penn World Table 8.0 (Feenstra et al., 2013), now provided by the Groningen Growth and Development Centre (GGDC), introduces a wide range of methodological changes, many with the aim of increasing consistency across versions and reducing the amount of speculation in the reported data, e.g. by removing 22 countries with particularly poor data coverage from the table. In the process, the authors also discarded the RGDPL series JMT base their analysis on. I therefore use the conceptually most similar series included in the dataset, labelled RGDPna, which the authors confirm to be the measure most in line with previous versions of the PWT (Feenstra et al., 2013). While it also corresponds to PPP adjusted GDP at 2005 prices, the growth rates here are taken directly from the underlying National Accounts data. It is therefore widely unaffected by some fundamental innovations in PWT8. It is also worth noting that the authors do not explicitly report the expenditure shares employed in the study at hand in the PWT release. They do, however, provide the underlying National Accounts data which contains this information. Note also that the reduced coverage means that PWT8 only covers 33 of the 36 countries covered by JMT for a sufficiently long period, leaving us with a slightly smaller sample.

The World Bank's World Development Indicators (The World Bank, 2015) provide two different series of GDP at constant prices - in 2005 US dollars, and in PPP adjusted 2005 US Dollars. While the second option is conceptually closer to the PWT measures JMT and I employ, it also has a much lower coverage of countries. Furthermore, the nature of the analysis only gives a minor role to absolute differences in income, as it does not exploit the cross-sectional dimension, but instead focusses on temporal variation. Furthermore, all series are in constant prices, implying that the temporal variation in prices is not reflected through the PPP adjustment in any of the series. Differences may, however, be induced through the relative valuation of individual expenditure share. For instance, a very low price for investment may substantially inflate the real GDP estimate in a period when investment is relatively high compared to the other measures. As the amplitude of the estimated coefficients (which would indeed vary with the levels) is not our primary interest, there is in the present analysis no inherent reason to use PPP adjusted series rather than to series valued at the exchange rate. I therefore opt against the PPP adjusted series, in order to retain an already dramatically reduced sample of 13 countries for the WDI.

All aid data are ODA net disbursements as reported by the Development Assistance Committee (DAC) and published by the OECD, which is the same source and measure JMT use. The data can sometimes be subject to minor revisions, which proved, however, not to have a substantial impact on the results as I will show in section 5.

3.2. Relative divergence of the datasets

To illustrate how these methodological discrepancies are reflected in the resulting data, figure 1 plots each of the GDP series for the countries that will be looked at more closely in the subsequent analysis (section 5.2.5), and figure 2 plots the corresponding investment shares, in order to illustrate differences in the reported composition of GDP. Anticipating the findings of section 5.1, the four countries at the top of the panel are those where JMT's results remain the most stable, the ones at the bottom are those with the least consistent results. The series are normalised to their respective 1965 levels in order to abstract from differences in levels that typically arise because of different underlying prices in the PPP estimates, and would mostly set WDI apart. Furthermore, as discussed earlier, these differences only play a minor part in the subsequent analysis.

Looking at figure 1, it seems that only in one country, Burkina Faso, GDP takes an almost identical trajectory in all four datasets, with only minor discrepancies over time.

In other countries, the discrepancies look rather well behaved, that is, they occur only at specific points in time or concern only a single series. Kenya is such a case, where WDI indicates that GDP has increased by a factor of about 6.5 from 1965 to 2007, whereas the PWT measures agree on a factor of about 4. The reverse is the case in Benin, where WDI indicates a persistently lower growth rate subject to similar fluctuations as that reported by the PWT measures. In Cameroon, all measures follow an almost identical path until the mid-80's, but split up: PWT8 and WDI register continued growth until the late 80's, followed by about a decade of recession, PWT6 and PWT7 start indicating a similarly severe recession earlier[3]. The resulting difference is preserved in subsequent periods, where the datasets generally agree on the growth rates, but at a now different level of GDP.

In Togo, Mauritania, Gabon and Lesotho the differences are perhaps the most striking and least tractable. While the general patterns tend to be the same (they agree on major booms and recessions), differences arise throughout the period without following an apparent pattern, leading the graphs to intersect sometimes multiple times, that is, none of the datasets systematically under- or over-reports growth, but the sign of their relative bias varies in time. The most striking single discrepancy may be the one between PWT6 and the remaining series in the mid-70's in Gabon. This is almost entirely explained by different underlying prices of investment, also reflected in figure 2, as we will see.

[3]Kobou et al. (2008, p. 558) attribute this dramatic economic decline to a combination of social tensions, an appreciating exchange rate, and external factors such as volatile commodity prices.

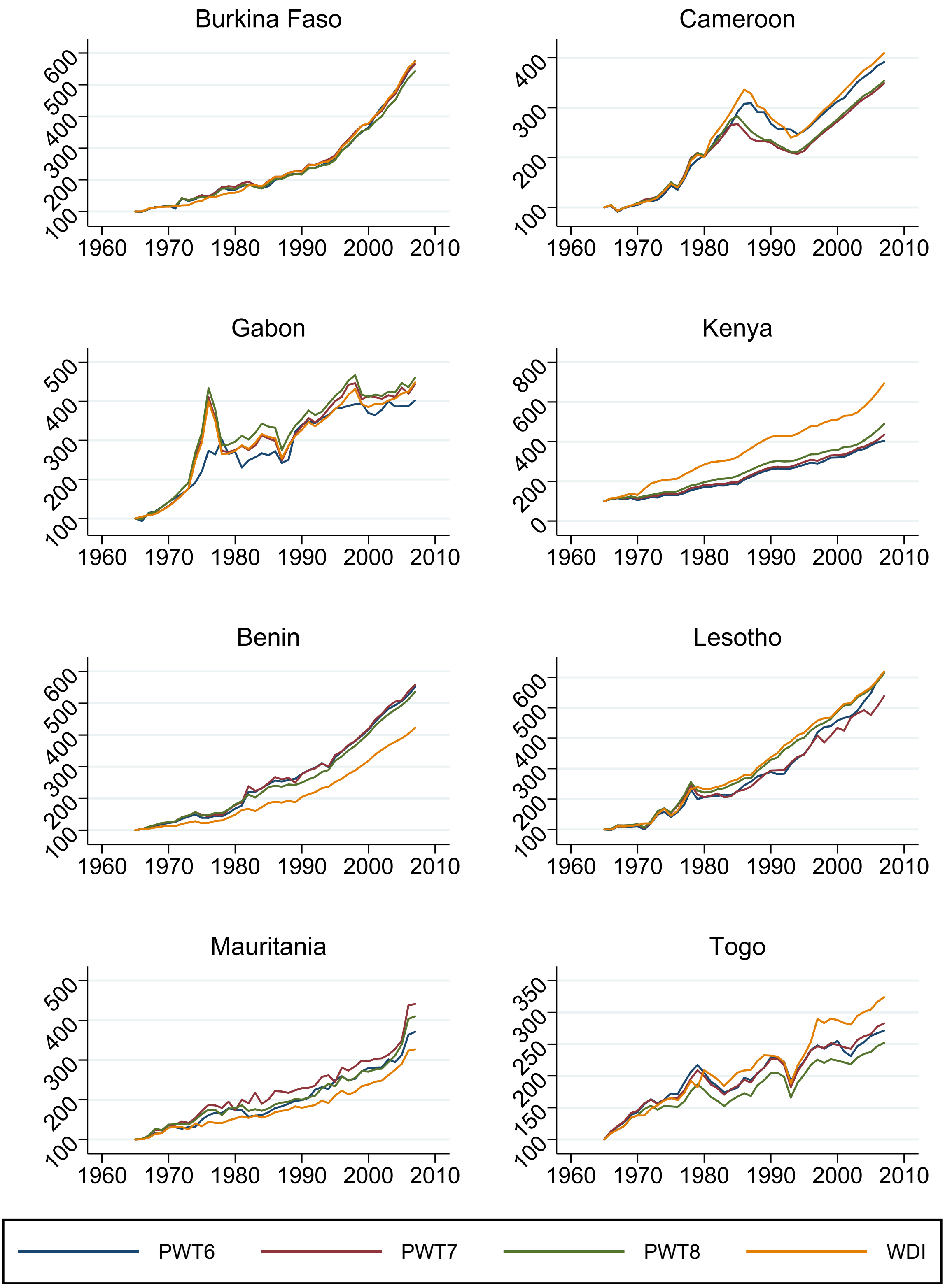

Figure 1: GDP series from 4 sources, normalised to 1965 levels

It is also worth emphasising that there is no obvious pattern describing how the datasets behave relative to each other: In Gabon, PWT6 is the obvious outlier, in Cameroon (and in Lesotho, to some extent), PWT6 and PWT7 take one path, PWT8 and WDI the other, in Togo, PWT8 and WDI are precisely the ones the diverge most relative to each other. Only Kenya and Benin are consistent with the perhaps most intuitive expectation, that is, WDI diverging from otherwise consistent PWT measures. However, even here the divergence has opposite signs.

Figure 2 plots the share of investment in GDP over the same period and for the same countries as 1 does for GDP itself. The first thing to note is that the y-axis is scaled in order to depict a maximum of detail in the variation for each country. Its range is therefore informative in itself, although it can vary with two factors: The temporal variation of the investment share within countries, and the discord between datasets regarding the investment share. For instance, investment in Lesotho varies from approximately 10% to more than 60% of GDP, but this is mainly due to temporal variation of similar amplitude in all datasets. In Cameroon, the scale is mainly stretched by upwards outliers in the 1980's in WDI (up to almost 40%), and a consistently much lower estimates of less than 10% in PWT6.

In most instances, the discrepancy is relatively constant over time, reflected in more or less parallel paths of the graphs; this applies in particular to Gabon, Kenya, and in a less pronounced manner to Burkina Faso, Togo, and Benin.

The obvious outlier in the panel is Mauritania: While the PWT series already diverge in a significant manner, this is dwarfed by the path suggested by WDI. Consistent with the other series until the early 1980's, the share of investment then skyrockets to levels of around 60% while the others agree on 10-20%, and reaches a peak of more than 150% of GDP (possible through an enormous trade deficit) where the PWT series report values

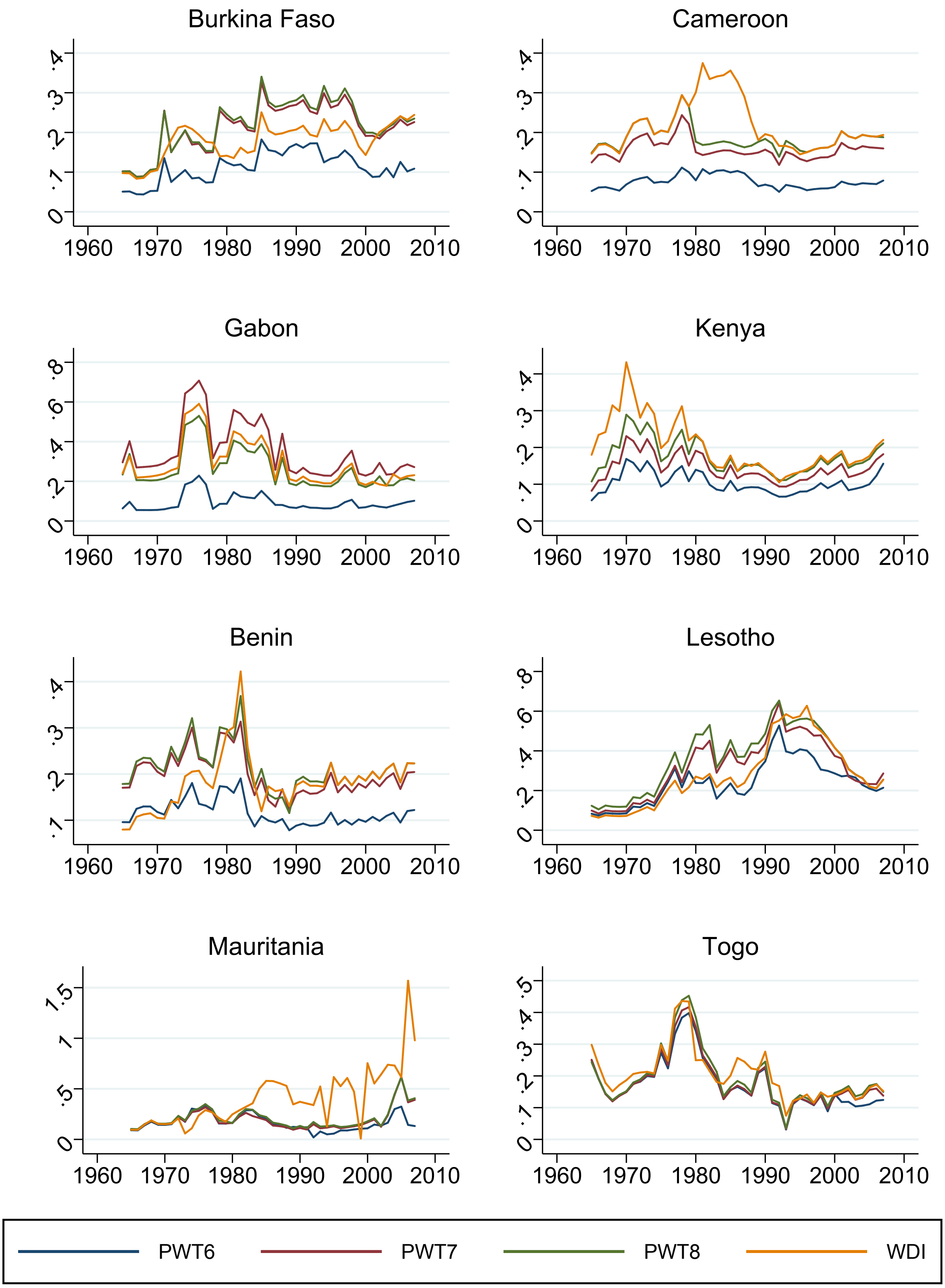

Figure 2: Shares of investment in GDP from 4 sources

between 20% and 40%.

Unlike in figure 1, one pattern across datasets is quite striking here: PWT6 almost consistently reports a lower share of GDP than the other datasets. One explanation that will be touched in section 3.1 is that PWT6 relies on price estimates from the ICP 1996, whereas the other series rely on those from 2005. A higher price for investment goods, mechanically reducing the calculated real share of investment if nominal figures remain the same, may either be a result of the differences in methods of the ICPs, or of different methods of processing them, or reflect an actual decline in these prices. Given that all of our countries experienced significant growth in the period between 1996 and 2005, the latter would be consistent with the finding that relative investment prices typically decrease with the level of development of an economy (e.g., Hsieh and Klenow, 2003).

		PWT6	PWT7	PWT8
Δ GDP	PWT7	0.76		
	PWT8	0.75	0.87	
	WDI	0.65	0.75	0.85
Investment	PWT7	0.54		
	PWT8	0.49	0.88	
	WDI	0.42	0.59	0.55
Consumption	PWT7	0.80		
	PWT8	0.76	0.85	
	WDI	0.66	0.77	0.74
Government	PWT7	0.71		
	PWT8	0.74	0.70	
	WDI	0.49	0.27	0.61

Notes: The reported values are pairwise correlations between each of the variables in the four datasets employed. Δ GDP are annual growth rates, the remaining variables annual shares of GDP.

Source: Author's calculations.

Table 1: Correlations between the core variables

Table 1 reports the pair-wise correlations between the annual GDP growth rate, and the shares of investment, consumption and government expenditure in the four datasets employed in JMT's sample of countries in the analysis over the period from 1965 to 2007[4].

Given that they are conceptually in principle the same, it is a striking result that the highest correlation of the growth rates between any of the datasets is 0.87 (PWT7 and PWT8), and goes as low as 0.65 (PWT6 and WDI). As the further analysis will focus on comparing results obtained from PWT6 with those obtained from each of the other measures, it is also important to note that for PWT8 and WDI, the lowest correlation is precisely that with PWT6, and PWT7 is only marginally less correlated with WDI than with PWT6.

The share of investment, consistent with what we tentatively inferred from figure 2, is only weekly correlated between PWT6 and the other measures, with coefficients around 0.5. Only PWT7 and PWT8 seem to generally agree over this variable, yielding a correlation coefficient of 0.88. This is, again, consistent with what we would expect from figure 2.

The remaining two measures, the share of consumption and government expenditure in GDP, seem to be more consistent across the datasets than investment. A notable exception is government expenditure in WDI, where the correlations go as low as 0.49 with PWT6, and 0.27 with PWT7.

A general and intuitive pattern in table 1 is that the correlations go down as we move away from the original dataset – PWT6, by these measures is more similar to PWT7 than to PWT8, and the least similar to WDI, which is both the most recent dataset and the one that belongs to a different series.

[4]Subject to data availability, see section 5.1.

4. The CVAR methodology

Consider a vector of $p = 5$ dependent variables $\mathbf{X}_t = [aid_t,\ GDP_t,\ inv_t,\ cons_t,\ gov_t]'$, where the variables correspond to the inflow of foreign aid, total GDP, investment, private consumption and government expenditure at time t respectively. This vector can be represented as an k-th order autoregressive process

$$\mathbf{X}_t = \mathbf{\Pi}_1 \mathbf{X}_{t-1} + \mathbf{\Pi}_2 \mathbf{X}_{t-2} + \mathbf{\Phi} \mathbf{D}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim iidN(0, \mathbf{\Omega}) \tag{1}$$

where $\mathbf{\Pi}_i$ are unrestricted matrices of autoregressive parameters. $\mathbf{D}_t$ is a $m \times 1$ vector containing m the deterministic components of the model, such as trends in the variables or dummies accounting for extraordinary events. These enter the model with the coefficients in $p \times m$ matrix $\mathbf{\Phi}$. $\boldsymbol{\varepsilon}_t$ is a $p \times 1$ vector of white noise residuals with constant variances and covariances embodied in the $k \times k$ matrix $\mathbf{\Omega}$. For the purpose of illustration, and because effectively this describes all cases in this study, the discussion here is restricted to the case of $k = 2$ lags.

The macroeconomic variables represented by $\mathbf{X}_t$ are typically found to exhibit non-stationary behaviour. If this is the case, and cointegration occurs between them, process (1) can be represented in its error correction form (ECM) following Engle and Granger (1987)'s representation theorem. Subtracting $\mathbf{X}_{t-1}$ from both sides, we get:

$$\Delta \mathbf{X}_t = \mathbf{\Pi}\, \mathbf{X}_{t-1} + \mathbf{\Gamma}_1\, \Delta \mathbf{X}_{t-1} + \mathbf{\Phi} \mathbf{D}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim iidN(0, \mathbf{\Omega}) \tag{2}$$

where $\mathbf{\Pi} = -(\mathbf{I}_p - \mathbf{\Pi}_1 - \mathbf{\Pi}_2)$ ($\mathbf{I}_p$ being a $p \times p$ identity matrix), and $\mathbf{\Gamma}_1 = -\mathbf{\Pi}_2$ is a $p \times p$ matrix of short-run adjustment coefficients. Δ is the first difference operator $(1 - L)$ with L defined such that $L\mathbf{X}_t = \mathbf{X}_{t-1}$.

If cointegration occurs between the variables in $\mathbf{X}_t$, $\mathbf{\Pi}$ has reduced rank $r < p$ and can be factorised such that $\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $p \times r$ matrices of full rank, economically interpretable as long-run equilibrium relations (see Johansen, 1996, chapter 3 and Juselius, 2006, chapter II.5).

Another corollary from the Engle-Granger representation theorem, of particular relevance in this context, is that equations (1) and (2) can be mathematically equivalently represented as a moving average (MA) process

$$\mathbf{X}_t = \mathbf{C}\sum_{i=1}^{t}\varepsilon_i + \mathbf{C}\,\Phi\sum_{i=1}^{t}\mathbf{D}_i + \mathbf{C}^*(L)\varepsilon_t + \mathbf{A}_0 \tag{3}$$

where $\mathbf{C} = \boldsymbol{\beta}'_\perp(\boldsymbol{\alpha}'_\perp\Gamma\boldsymbol{\beta}'_\perp)^{-1}\boldsymbol{\alpha}'_\perp$ is a $p \times p$ matrix of rank $p-r$, and $\boldsymbol{\alpha}_\perp$ and $\boldsymbol{\beta}_\perp$ are the $p\times(p-r)$ orthogonal complements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that $\boldsymbol{\alpha}'\boldsymbol{\alpha}_\perp = \boldsymbol{\beta}'\boldsymbol{\beta}_\perp = 0$. $\mathbf{C}^*(L)$ is a stationary lag polynomial and $\mathbf{A}_0$ contains the initial values of the variables in $\mathbf{X}_t$ and the initial values of the short-term dynamics.

The main parameters of interest in this context are contained in the long-run impact matrix $\mathbf{C}$, which has the following structure:

$$\mathbf{C} \quad = \quad \begin{array}{c c} & \begin{array}{ccccc} \hat{\varepsilon}_{aid} & \hat{\varepsilon}_{y} & \hat{\varepsilon}_{inv} & \hat{\varepsilon}_{c} & \hat{\varepsilon}_{g} \end{array} \\ \begin{array}{c} aid_t \\ y_t \\ inv_t \\ c_t \\ g_t \end{array} & \begin{pmatrix} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} \\ c_{21} & c_{22} & c_{23} & c_{24} & c_{25} \\ c_{31} & c_{32} & c_{33} & c_{34} & c_{35} \\ c_{41} & c_{42} & c_{43} & c_{44} & c_{45} \\ c_{51} & c_{52} & c_{53} & c_{54} & c_{55} \end{pmatrix} \end{array} = \begin{pmatrix} c_{11} & \mathbf{C_{12}} \\ \mathbf{C_{21}} & \mathbf{C_{22}} \end{pmatrix} \tag{4}$$

The element c_{ij} in row i and column j shows the impact that the cumulated exogenous shocks, measured as the residuals of the respective equation, on variable j have exerted on variable i in the long run. In this sense, for example, c_{21} can be interpreted as the long run effect of exogenous shocks to foreign aid on GDP.

It follows from the research interest of JMT and the study at hand - the long-run impact of foreign aid - that the main focus of the discussion will lie on the elements in column 1, reflecting precisely that. In line with JMT, equation (4) provides a convenient notation of the matrix $\mathbf{C}$, where the four coefficients describing the long-run impact of aid on the other variables is contained in $\mathbf{C}_{21}$, while the inverse effect - the long-run impact of the other macrovariables on aid - is described by $\mathbf{C}_{21}$. The latter is informative when asking questions about the potential (and plausible) endogeneity of aid. The $(p-1) \times (p-1) = 4 \times 4$ matrix $\mathbf{C}_{22}$ contains the remaining parameters, having analogous interpretations for the interaction between the macrovariables variables other than aid.

5. Assessing the impact of the data

5.1. Simple replication

The first replication uses the models developed in the original study, reported in table 2, JMT. This involves a replication using the original data (PWT6, except for Sudan where JMT use WDI data) for all 36 countries, and the same exercise employing the alternative datasets. JMT's mode of inference of considering the first best and second best choice of rank requires two estimations per country and dataset (one for each rank). Besides the original data for the 36 countries in JMT's sample, we have sufficient data for 36 (PWT7), 33 (PWT8), and 13 (WDI) of these countries in the alternative datasets, each of which has to be estimated twice, leaving us with 236 country-, dataset-, and rank-specific estimations. These have been carried out in an appropriately modified version of CATS in RATS version 2.6[5].

[5]Replication programs and data can be obtained from the author upon request. Parts of the relevant code are embedded in proprietary program files and can therefore not readily be shared, but detailed documentation on the necessary modifications can be provided.

5.1.1. Criteria to assess the stability of the results

When assessing the stability of JMT's results with respect to changes in the data, given that their mode of inference is rather unconventional, it is necessary to establish some sensible and objective criteria. Their answer to the main question, the long-run impact of foreign aid on economic development, is based on the sign and magnitude of the t-ratios of the coefficients of the **C**-matrix, discussed in section 4. Specifically, they focus on the coefficients that describe foreign aid's impact on GDP (Y) and Investment (I), but also report those on Private Consumption (C) and Government Consumption (G). Aid, in this context, is defined as effective if either the coefficient of I, or Y, is positive and significant, or if both are – significance being defined as having a t-ratio larger than 2. While the representation of the results also highlights marginally insignificant coefficients (between 1.6 and 2), these do not impact on the final inference. Aid harmfulness is defined analogously, with opposite signs.

My discussion will evolve around the following criteria:

1. The number of reversals in *any* of the coefficients, i.e. Y, I, C and G. While JMT do not base their overall inference on C and G, they are still reported and discussed - again, counting positive coefficients as evidence in favour of aid effectiveness.

2. The number of reversals in the *most relevant coefficients*, Y and I.

3. The number of reversals in the *inference by country*. This is not equivalent to the previous point, as the inference is a joint product of the two coefficients of Y and I - even if both change sign/significance, we would still infer effectiveness as long as one of them is positive significant.

4. Any changes in the *overall conclusion*. JMT count 27 out of 36 cases of aid effectiveness. Irrespective of the results obtained for the previous criteria, this

ratio may change or remain approximately constant, as country-specific reversals may be compensated for by reversals in the opposite direction for other countries.

For criteria 1 and 2 (coefficient specific), I consider changes in the coefficients a *reversal* if their absolute t-value changes at least from $|t| > 2$ to $|t| < 1.6$ (looses significance), does the opposite (gains significance), or if it changes signs without remaining insignificant in both instances (that is, in JMT's original study and in my replication).

For criterion 3, the definition of a reversal is more straightforward. The inference by country is well defined in the original paper and takes three possible values (effectiveness, harmfulness, or neither), and I consider any change in this value a reversal.

Overall inference (4) boils down to a simple ratio between the number of cases of effectiveness and harmfulness respectively, and does not require further definition.

Note also that JMT's approach of adopting different economic priors adds a further complication. Instead of reporting the results only from the preferred specification, their main focus is on the results choosen from the best and the second best cointegration rank, determined by a procedure that will be discussed in more detail in section 5.2.4 (summarised results are also reported for strictly the preferred rank, and first, second and third best choice of rank, and their results are robust to these different search algorithms). Under the prior of aid effectiveness, they report the preferred choice of rank if it indicates effectiveness, but the second if the first doesn't but the second does. The same is done under the prior of harmfulness, but looking for negative coefficients.

This practice does indeed provide transparency, as well as a sensible robustness check. In the study at hand however, it means that cases can arise where, while the inference for a country (criterion 3) does not change, it will be based on a different rank, with the originally reported rank now indicating a different result. The strictest approach would be to count such cases as inconsistencies, even if the inference is eventually the same. Besides the fact that this issue almost never arises, it should also be noted that,

	Replication	Alternative datasets		
	PWT6	PWT7	PWT8	WDI
Same inference	97%	67%	61%	77%
Consistent coefficients	88%	63%	58%	63%
Reversed coefficients	5%	28%	26%	23%
Effectiveness	26/36	18/36	13/33	6/13
Harmfulness	10/36	9/36	7/33	3/13
Insignificant	8 (18)	17 (20)	17 (21)	7 (7)
Sample	36	36	33	13

Notes: The first row reports the number of countries that yield the same inference as JMT, the second and third report the number of coefficients that are in line with their results. Effectiveness and harmfulness are counted under the respective priors. The number of insignificant coefficients is reported under the prior of effectiveness, followed by the figure obtained under the opposite prior in parentheses.

Source: Authors calculations.

Table 2: Summary of replication results

where it does, the same procedure still yields the same inference. Furthermore, our main interest remains the overall conclusion of aid effectiveness. By a law of large numbers type of argument, assuming that the conclusion of aid effectiveness in this framework is not systematically dependent on methodological choices underlying our datasets (e.g., a certain rank tends to favour the conclusion of aid effectiveness), it may well turn out that while the inference changes in a number of individual countries, these changes are of different signs, and the overall result remains consistent. I will therefore opt for the more conservative approach and count all cases where the inference remains the same as consistent.

In what follows, I will explore criteria 1 to 4, and thereby assess the robustness of the results both by country and by dataset in several dimensions.

5.1.2. Replication of JMT using Penn World Table 6.3

This section describes the results obtained from applying JMT's specifications, as reported in table 2 of their paper[6], to the same data they employed. This is PWT6 in all cases except Sudan, where JMT recur to WDI for reasons of data availability (WDI goes back to 1960, PWT6 only to 1970). Differences in the results may emerge because of revisions in the ODA data by OECDStats. In the case of Sudan, it can also not be guaranteed that the data employed by JMT are identical to mine, as I use a more recent version of WDI. However, the replication results are very close in this case, so this is not a major concern.

Under the economic prior of aid effectiveness, out of 144 coefficients, 127 (88%) indicate the same sign and significance as reported by JMT, and another 10 only change in a minor way, e.g., from marginally significant to insignificant, or from positive and significant to positive and marginally significant. 'Reversals', defined as above only occur 7 times out of 144 (5%), only 2 of which concern our main coefficients of interest, those of I and Y.

Looking at changes in inference, the replicated results only differ in one case, Mauritania, where the coefficient of I turns from positive significant to negative significant, reversing the inference. This almost certainly reflects significant revisions to the ODA data. Overall, the inference of JMT under the prior of aid effectiveness remains the same, with now 26 instead of 27 out of 36 cases indicating significant positive effects of aid on long-term growth, 2 negative, and 8 none. Under the prior of aid harmfulness, my results are identical to JMT's, with 6 instances of significant negative effects on GDP and 5 on investment, only one of which overlap, leading to indications of harmfulness in 10 cases.

[6] In two cases, Nigeria and Zambia, the reported sample periods are different from those appearing in the output provided by the authors (documenting rank specification). The latter are more in line with the reported results, and in the case of Nigeria one reported dummy lies outside the reported sample period. I therefore adopt the models implied by the output.

My replications are thus very close to JMT's original results, and the revisions on the DAC ODA data reported by the OECD do not have a major impact on the results. However, I note as a caveat that in 6 countries, namely Botswana, Mauritania, Rwanda, Sudan, Seychelles and Togo, individual coefficients have been significantly influenced by the changes in the aid data. This only applies to a single coefficient in all countries but one, and leads to different inference only in the case of Mauritania. We shall bear this in mind when extending the analysis to further datasets.

5.1.3. Robustness check using Penn World Table 7.1

As pointed out in section 3, PWT7 differs from PWT6 in two key aspects, the base year, and the concepts of household and government consumption.

Out of our 144 coefficients (PWT7 covers all 36 countries in the sample), only 89 (67%) remain of the same sign and significance, 15 change in a minor way, and in 40 cases (28%), significant changes occur to the results. About half of the latter category (21 coefficients) concern one of the key variables, investment or GDP.

The impact on the country-wise inference is sizeable, but less pronounced. It remains identical in 24 cases, is reversed in one case (from positive significant to negative significant in Lesotho), and loses or gains significance 11 times.

Overall, the conclusion of aid effectiveness is still reasonably well backed by the results, with 18 out of 36 cases in favour of aid effectiveness when assessed under this prior, and only 9 indicating harmfulness when assessed under the corresponding prior.

5.1.4. Robustness check using the Penn World Table 8.0

In an attempt to reduce the guesswork underlying the numbers reported in the PWT, the new authors behind PWT8, besides returning to the more classical concept of household consumption, reduced the coverage of countries. In our sample, this affects Ethiopia, Somalia and the Seychelles, leaving us with a sub-sample of 33 countries.

Out of the 132 estimated coefficients, 76 (58%) remain of the same sign and significance as in JMT's analysis, 34 (26%) change in a significant manner, and the remaining 22 change in a less substantial way.

In terms of country-wise inference (under the prior of aid effectiveness), the replication using PWT8 agrees with JMT in 20 countries (61%), disagrees in 11, and yields opposite inference in 2 countries, Liberia and Lesotho.

The support for the original conclusion of aid effectiveness remains, but again becomes less clear with these data. Under the prior of aid effectiveness, effectiveness can be discerned in 13 countries, harmfulness in one, and either of the hypotheses lacks support in a majority of 17 countries. When assessed under the corresponding prior, the hypothesis of harmfulness finds support in 7 countries. Effectiveness is then only supported in 5 countries, and the vast majority of 21 cases gives no support to either of the hypotheses.

5.1.5. Robustness check using the World Development Indicators (2015)

Our sample is dramatically reduced to 13 countries when using WDI, due to a significantly lower coverage of countries (countries with low statistical capacity in particular). Effectively, new information only comes in for 12 countries, as JMT used WDI for Sudan, which I therefore did myself for the direct replication. For the sake of completeness, I will include it here anyway.

Under the prior of aid effectiveness, out of the remaining $13 * 4 = 52$ coefficients, 33 or (63%) are in line with those obtained by JMT, 12 (23%) change in a significant manner, and another 7 change in minor ways.

The conclusion however remains constant for 10 out of 13 (77%) countries, with now 6 countries indicating effectiveness, 7 neutrality (or nothing), and no evidence for harmfulness. Under the prior of harmfulness, support for the latter can be found in 3 countries.

It should be noted that, while these results are proportionally speaking more in line with JMT than those obtained with PWT7 and PWT8, the significantly smaller sub-sample contained in WDI consists of countries that exhibit more stable results in the other datasets. In the sub-sample of countries that are not covered in WDI, PWT7 and PWT8 yield consistent results only in 24 out of 43 cases (56%), while they yield consistent results in 20 out of 26 cases (77%) in the sub-sample covered by WDI[7]. This could possibly be reflecting a selection bias as a function of the quality of the original data, if for instance WDI are more conservative in "filling the gaps".

5.1.6. Consistency across countries

The number of consistent coefficients, excluding the PWT6 replication, ranges from 0% in Botswana and Lesotho, to 100% in Chad, Burkina Faso, Cameroon, Kenya and Gabon. The median proportion of consistent coefficients is 50%, the mean 57%. In Botswana and Lesotho all coefficients change in every dataset, although it should be noted that Botswana is one of the countries with a relatively poor initial replication (i.e., using PWT6, implying that the revisions to the aid data play an important role here). In five countries (Chad, Burkina Faso, Cameroon, Gabon and Kenya), all coefficients have the same sign and significance as in JMT. The table summarising the consistency by coefficients for each country has been relegated to appendix A.

Abstracting from the divergence in individual coefficients and focussing on the country-wise inference of aid effectiveness shows that almost half the sample, 17 countries, yield the exact same inference throughout all datasets. On the other hand, 9 countries do not yield the same inference with any of the new data. Only in 3 out of a total of 82 cases, the inference is reversed. These are Lesotho in PWT7 and PWT8, and Liberia in PWT8; all three reversals correspond to a switch from effectiveness to harmfulness. Virtually all other changes in inference correspond to a loss of significance, therefore supporting either aid ineffectiveness or none of the hypotheses, depending on one's interpretation

[7]The fact that this corresponds exactly to the ratio found in the replication using WDI is coincidental; only half of the reversed inferences are associated with the same countries as in WDI.

of the absence of statistical significance.

5.2. Re-specification for selected countries

The second exercise aims at exploring the full potential impact of the data onto the results of JMT and similar studies. It takes into account a fundamental component of the philosophy underlying the CVAR approach, namely that of "allowing the data to speak freely" (Hoover et al., 2008), meaning that the modelling in this approach is mainly inspired by the data, with the priority of creating a statistically adequate representation of the latter before making strong theoretical assumptions about the mechanisms at work. Each country-specific model employed in JMT and thus adopted in section 5.1 has, in this spirit, been specified as a function of the underlying data, that is PWT6 and the DAC ODA data.

Given the sometimes striking divergence between the datasets, discussed in section 3 and shown to potentially have a sometimes impact on the inference in section 5.1, it seems reasonable to expect that differences may emerge not only in the results, but, on a more fundamental level, in the models based on these different data.

I will therefore re-specify the models for a sub-sample of selected countries. In order to explore both ends of the spectrum, my analysis focusses on the four most consistent and inconsistent cases respectively, conditional on sufficient data availability in all datasets. Consistency is here quantified by the total share of coefficients that are consistent with those obtained by JMT. Namely, the most consistent countries meeting this criterion are Burkina Faso, Cameroon, Gabon and Kenya, the least consistent ones are Benin, Mauritania, Lesotho and Togo.

The models employed in JMT have the following variable elements: deterministic components such as trends and dummies (section 5.2.1), the lag-length k (section 5.2.2), and the cointegration rank (section 5.2.4). While the econometric literature offers a plethora of formal criteria to specify either of these components, it needs to be emphasised that the sample size in this application ($T \approx 40$) undermines the power of some of the relevant tests. Furthermore, different criteria typically offer a slightly different angle at the data, and will sometimes indicate different choices. The resulting trade-offs will eventually need to be resolved by the researchers judgement. The following section aims at establishing a rigorous specification procedure, reflecting JMT's approach as closely as possible, and apply it to the aforementioned countries for the four datasets employed.

5.2.1. Deterministic components

Linear trends The choice of the deterministic components in this model essentially consists out of two elements. The first is that, given the macroeconomic nature of our data, we would typically expect the presence of linear trends. These can be incorporated in the VAR in different ways, and the decision essentially boils down to the question whether the trends cancel in the cointegrating relations or not. In the case where they do cancel out, the most accurate specification would include an unrestricted constant term in the equation. If they do not cancel out, a trend that is restricted to the cointegrating relations should be included; this is the case when either some of the variables in $\mathbf{X}_t$ or any of the cointegrating relations are trend-stationary. For a detailed discussion, see Juselius (2006, chapter 6). One way of approaching the issue is to tentatively include a trend that is restricted to the cointegrating relations, which can then be tested for significance. In all of the present cases, this evidence unambiguously points towards the inclusion of a trend restricted to the cointegrating relations (see Juselius (2006, p. 100), Case 4).

Dummy variables The more contentious, and less clearly defined choice, concerns the inclusion of dummy variables in order to account for extraordinary events such as droughts, floods, social unrest, or changes in equilibrium relations due to, e.g., regime changes. As it is difficult *a priori* to determine which historical events have an impact significant enough enter the model, Juselius (2006, chapter 6.6) suggests to first scrutinize the data and the residuals from the baseline VAR, in order to determine where it is required to correct for outlier observations; the modelling does thus first depend on the statistical evidence, which is then complemented by institutional knowledge.

In line with JMT and Juselius (2006), I use three classes of dummies: Permanent blip dummies, labelled D_pZZ_t, having the structure $[0,0,1,0,...,0]$, transitory blip dummies $D_{tr}ZZ_t$ with the structure $[0,0,1,-1,0,...,0]$, and shift dummies D_sZZ_t, restricted to the cointegrating relations and taking the form $[0,0,1,1,1,...,1]$, indicating a shift in the equilibrium mean of the cointegrating relations.

From a data perspective, the indications for the inclusion of a dummy are primarily derived from an inspection of the residuals. Large residuals, "large" here being construed as corresponding to approximately 3 standard deviations, indicate that there may be an extraordinary event which, if not accounted for appropriately, would distort the analysis[8]. A unique blip in the error series, e.g., if it is not reverted by a shock in the opposite direction in one of the following periods, may then indicate a unique event (e.g., a drought) that permanently affected the economy. A temporary blip, for example, a large positive residual followed by a large negative residual in the following period, is an indication for a transitory intervention; typical cases would include a period of expansive monetary policy, compensated for by contractionary policy later on, or temporary fiscal stimuli.

The determination of a shift dummy, D_sZZ, is in practice less straightforward, as it accounts for level changes in the long-run equilibrium, which are not readily observable from looking at the residuals. However, graphical inspection of the cointegrating rela-

[8]The outliers detected this way were generally consistent with those discerned through the dummy saturation procedure implemented in Autometrics (see Doornik, 2009). Due to the fundamental importance of institutional knowledge the present set-up, careful inspection of the residual series generally served as the primary source of information.

tions can provide some indication. Furthermore, Juselius (2006, chapter III.9) proposes a battery of recursive and backward recursive tests that can provide indications for shifts in the equilibrium relations, which are taken into consideration.

For the final choice of the dummy variables, the statistical evidence is further complemented with historical data from the UCDP/PRIO Armed Conflict Dataset (Pettersson and Wallensteen, 2015) and the EM-DAT International Disaster Database (Guha-Sapir et al., 2014)[9], as well as knowledge about economically relevant historical episodes, most of which are documented in Ndulu et al. (2008).

Figure 3 depicts the resulting dummy variables on timelines corresponding to the respective sample periods for each country and dataset, where the top line for each country corresponds to the model for the original PWT6 data, the second to PWT7, the third to PWT8 and, and the fourth to WDI. The presence of apparent bulks of dummy variables within each country across datasets illustrates the fact that the datasets tend to agree on the impact of major events. Take for instance $Dp83_t$ in Lesotho, the year of a severe drought affecting most of the population(Guha-Sapir et al., 2014), reflected here in a dramatic drop in investment. Another consistent outlier is $Dtr7478_t$ in Gabon, corresponding to a unusually large temporary increase in investment, likely spurred by sharp increases in the price of oil after the 1973 oil crisis. Or the $Ds94_t$ shift dummy in Kenya, which coincides with a severe drought coupled with a dysentery epidemic costing about 1000 lives (Guha-Sapir et al., 2014).

The latter also illustrates the trade-off one faces when weighing institutional knowledge against the statistical evidence. While the historical events are themselves not necessarily susceptible to induce a shift in long-run equilibrium relationships, the statistical evidence

[9]The events documented in these two databases have been matched with the dummies included in all models in a table available at `http://bit.ly/1W2IVQe`.

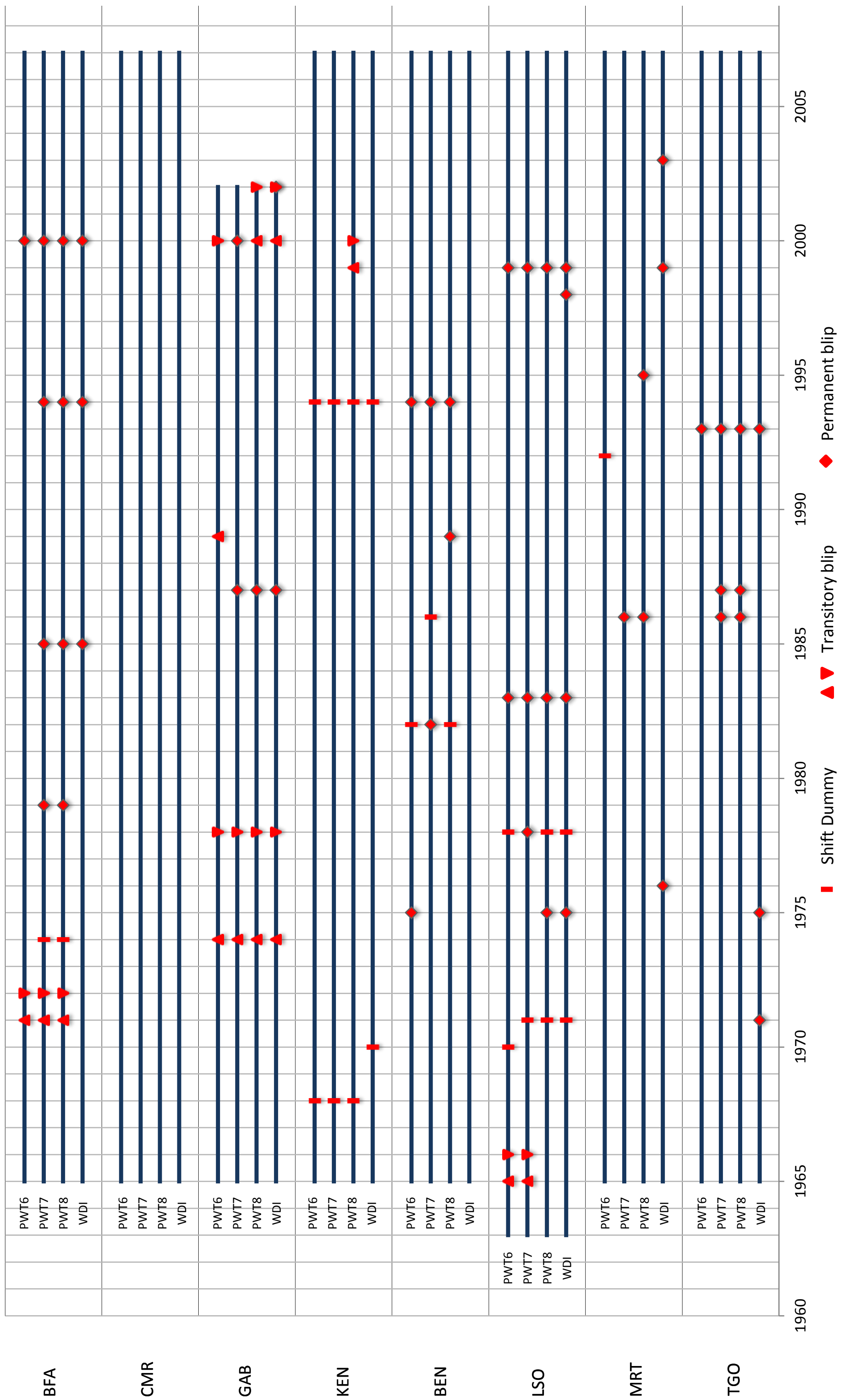

Figure 3: Timeline of dummy variables

suggests precisely this, with the recursive test for fluctuations of the eigenvalue indicating non-constancy of our α and β parameters, with a sharp decline in the test statistic in 1994 for the third cointegrating relation. The inclusion of $Ds94_t$ establishes constancy over the entire sample period, justifying the inclusion both in JMT's and in my models.

On the other hand, there are about 15 dummy variables that are specific to the respective dataset, such as in Togo in 1971 (a decline in GDP and investment in the year of a drought affecting about 150.000 people (Guha-Sapir et al., 2014)) or 1975/76 (a spike in consumption and government spending coinciding with soaring phosphate prices, one of Togo's staple exports (Ndulu et al., 2008, Vol. 1, p. 150)). Particularly divergent, in terms of the modelling of extraordinary events, is Mauritania, where the 7 dummies included across the models occur in 6 different years, thus only agreeing once. Nevertheless, all of them can plausibly be attributed to historical events (1976 is the year of a severe drought effectively affecting the entire population; 1986 falls within a period of severe ethnic tensions and shortly follows a coup d'état by Ould Taya; 1992, a shift in equilibrium means, is the year of the adoption of a new constitution establishing the fourth republic; 1995, 1999 and 2004 have all seen major floods (Guha-Sapir et al., 2014; Seely, 2005)).

Note that some of the peculiarities in the data can not that clearly be attributed to historical events, in which case the trade-off is between the statistical fit of the model, and its institutional/historical justifiability. The 1971/72 transitory dummy in Burkina Faso, indicated in all datasets except WDI, is such a case. While this falls within a period of some political instability, with a tightening of military control and a decline of the civilian role (Ndulu et al., 2008, Vol. 2, p.8 Appendix), there is no compelling historical evidence justifying the transitory nature of the shock. It is, however, by some margin the specification that best fits the pattern of the residuals, and my models therefore follow JMT in accounting for it with $Dtr71_t$. In other cases, the trade-off has been resolved in favour of institutional adequacy, for instance in the case of an outlier in 1978 in the PWT7 Cameroon model: The model achieves a reasonable fit overall without a correction, and I could not identify a strong case for an extraordinary event in the country's history.

Overall, JMT's and my models (consistent in PWT6 except for Burkina Faso, where the aid data appears to have significantly changed the requirements) follow the same patterns and account for the statistical and historical evidence in a coherent manner. To summarise, the data tends to tell a similar story – stylised in figure 3 – but puts varying emphasis on individual episodes. The relevance of these nuances in intonation is the subject of this investigation.

5.2.2. Choice of lag-length

Determining the optimal lag-length of the resulting VARs boils down to a trade-off between preserving a maximum of information, and retaining a reasonably small number of parameters, especially given the relatively small number of observation for each country (*small* in the context of a Time Series Analysis). My choices are based on three groups of criteria, in line with Juselius (2006): (i) a likelihood test for lag reduction, (ii) information criteria, namely the Schwarz Criterion (SC) and the Hannan-Quinn Criterion (HQC), and (iii) a Lagrange-Multiplier test for no autocorrelation of the residuals. The p-values associated with (i) and (iii), as well as the values of the (ii) ICs are collected in table 3.

The (i) test for lag reduction tests for a significant decrease in the (log-)likelihood when adopting a more parsimonious model, which indicates a poorer fit of the latter. The first column of table 3 reports the outcome only of the test for the reduction to the final choice of lag-length k^* from k^*+1 (effectively from VAR(2) to VAR(1) in all cases) in order to save space. Note that this is rejected in almost all cases, including those exactly replicating JMT's models (all PWT6 except Burkina Faso), indicating that the second

		Schwarz		Hannan-Quinn		Autocorrelation			
Country / Data	*LR*	k^*	k^*+1	k^*	k^*+1	$k^*, 1$	$k^*+1, 1$	$k^*+1, 2$	k^*
Burkina Faso									
PWT6	0.00	-22.44	-21.39	-23.66	-23.28	0.47	0.74	0.39	1
PWT7	0.00	-24.57	-23.65	-26.46	-26.21	0.98	0.91	0.14	1
PWT8	0.00	-24.42	-23.46	-26.31	-26.02	0.97	0.85	0.09	1
WDI	0.00	-22.72	-21.87	-24.07	-23.89	0.24	0.02	0.41	1
Cameroon									
PWT6	0.12	-23.57	-22.10	-24.51	-23.72	0.57	0.75	0.60	1
PWT7	0.00	-24.07	-22.94	-25.01	-24.56	0.31	0.17	0.02	1
PWT8	0.00	-23.67	-22.61	-24.62	-24.23	0.17	0.26	0.03	1
WDI	0.90	-22.10	-20.21	-23.04	-21.82	0.97	0.08	0.75	1
Gabon									
PWT6	0.00	-16.83	-15.45	-18.43	-18.07	0.40	0.42	0.17	1
PWT7	0.00	-17.72	-16.20	-19.61	-19.25	0.56	0.72	0.21	1
PWT8	0.00	17.92	-16.45	-19.81	-19.51	0.56	0.53	0.13	1
WDI	0.00	-17.88	-16.50	-19.77	-19.55	0.43	0.54	0.34	1
Kenya									
PWT6	0.00	-23.88	-23.21	-25.36	-25.36	0.15	0.61	0.76	1
PWT7	0.00	-24.82	-24.21	-26.30	-26.37	0.09	0.52	0.32	1
PWT8	0.00	-24.95	-24.18	-26.57	-26.47	0.20	0.04	0.85	1
WDI	0.00	-24.88	-24.33	-26.36	-26.48	0.01	0.12	0.37	1
Benin									
PWT6	0.07	-25.78	-24.39	-27.26	-26.54	0.36	0.41	0.71	1
PWT7	0.00	-25.02	-23.75	-26.50	-25.91	0.34	0.34	0.14	1
PWT8	0.01	-24.58	-23.35	-26.06	-25.51	0.85	0.75	0.37	1
WDI	0.00	-24.12	-22.88	-25.06	-24.50	0.15	0.03	0.03	1
Lesotho									
PWT6	0.00	-20.49	-19.80	-22.32	-22.29	0.52	0.16	0.48	1
PWT7	0.00	-21.50	-20.58	-23.20	-22.93	0.84	0.77	0.83	1
PWT8	0.00	-22.85	-22.13	-24.69	-24.62	0.42	0.06	0.08	1
WDI	0.02	-20.79	-19.57	-22.23	-21.66	0.82	0.23	0.94	1
Mauritania									
PWT6	0.00	-17.80	-16.84	-19.01	-18.73	0.04	0.37	0.28	1
PWT7	0.11	-18.05	-16.60	-19.13	-18.35	0.74	0.21	0.57	1
PWT8	0.01	-17.76	-16.53	-18.98	-18.42	0.71	0.48	0.99	1
WDI	0.07	-16.96	-15.55	-18.31	-17.57	0.30	0.49	0.70	1
Togo									
PWT6	0.00	-21.42	-20.47	-22.50	-22.23	0.06	0.47	0.34	1
PWT7	0.00	-21.24	-20.53	-22.59	-22.55	0.13	0.91	0.35	1
PWT8	0.00	-21.02	-20.33	-22.37	-22.35	0.13	0.80	0.31	1
WDI	0.02	-19.21	-17.97	-20.56	-19.99	0.08	0.83	0.43	1

Notes: The reported values are p-values with the exception of the Schwarz Criterion and the Hannan-Quinn Criterion. k^* is the eventually inferred optimal lag length.

Source: Author's calculations.

Table 3: Criteria for the choice of the lag-length

lag may still contain valuable information. Although not explicitly reported here, this test rejects almost any reduction up to at least the fourth lag, which was the highest lag included in the procedure, therefore indicating prohibitively high autoregressive orders given the relatively small sample sizes we are facing. This is because it does not account for the fact that each new lag increases the number of parameters by p^2, thereby rapidly consuming the already scarce degrees of freedom.

The (ii) information criteria, explicitly designed to take this trade-off into account, will therefore tend to indicate shorter lag-lengths. Columns 3 to 6 of table 3 report the values of the SC and the HQC for both k^* and $k^* + 1$; the lag length associated with the smaller value is considered more favourable by the respective criterion. In all cases but four, the ICs unambiguously that indicate $k^* = 1$ is preferable over $k^* + 1 = 2$. In Gabon, PWT8, SC favours $k = 2$, while HQC favours $k = 1$, and in Kenya, the opposite is the case in PWT6, PWT7 and WDI. In these ambiguous cases, I opt for the more parsimonious specification. Note that this also seems to be the way JMT proceed in Kenya and PWT6. The fact that the ICs almost unambiguously indicate a short lag length is not surprising; both SC and HQC penalise the inclusion of new parameters more harshly, the shorter the T, and will therefore apply particularly high penalties in the short samples at hand.

Finally, columns 7-9 report the p-values from the (iii) LM test for no residual autocorrelation in the first lag of the VAR($k^* = 1$) models, and in the first and second lag of the VAR($k^* + 1 = 2$) models. The null hypothesis of no autocorrelation is rejected at the 5% level only twice in the first lag in the VAR($k^* = 1$) models (Kenya WDI and Mauritania PWT6). It does not, however, oppose $k = 2$ either in most cases, indicating residual autocorrelation only three times in the first lag and in the second lag respectively.

Overall, all datasets appear to be best described with a lag-length of $k = 1$. The most contentious choice is arguably Kenya in WDI, where HQC favours a higher lag-length, and there is evidence for residual autocorrelation in the first lag. In the light of otherwise reasonable properties of the model (discussed in the next section), I stick to $k = 1$ as indicated by SC, but note it as a possible caveat when interpreting the results.

5.2.3. Tests for Misspecification

Before proceeding to the choice of the cointegration rank, it is worth checking the models for misspecification, as not only the trace test, but the VAR model as such, rely on a number of assumptions - most crucially, that of normal and independent residuals. Table 4 reports the p-values for tests for autocorrelation in the second lag in the final model (the first lag having been considered and reported as a criterion for the choice of the lag length before, see table 3), multivariate normality, autoregressive conditional and heteroskedasticity (ARCH) in the first and second lag. The last column reports the trace correlation, roughly interpretable as an average R^2 over the five equations in the model, thus summarising its overall fit.

Note that there is no indication for residual autocorrelation in the second lag in any of the models when applying a 5% significance level, and only in two cases (Gabon PWT7, Benin PWT8) the test would reject at the 10% level (first column, table 4). Coupled with the overall good results with respect to the first lag reported in table 3, residual autocorrelation does not appear to be a typical problem in the models.

The second column reports the p-values resulting for the test of multivariate normality suggested by Hansen and Doornik (1994), based on a transformation of the relevant moments (kurtosis and skewness) proposed by Shenton and Bowman (1977), crucially adding to the validity of the test in the present set-up by enhancing its small sample properties. In the final models, multivariate normality is almost never rejected at the 5% level. The exceptions are Kenya PWT8, and Lesotho in all datasets but PWT8, which also includes the PWT6 model, identical to JMT's and yielding output almost

Country / Data	*Autocorr.*	*Norm.*	ARCH(1)	ARCH(2)	*Trace Corr.*
Burkina Faso					
PWT6	0.35	0.42	0.05	0.02	0.40
PWT7	0.37	0.11	0.32	0.07	0.63
PWT8	0.42	0.11	0.21	0.05	0.62
WDI	0.72	0.59	0.10	0.08	0.51
Cameroon					
PWT6	0.59	0.32	0.20	0.01	0.32
PWT7	0.47	0.28	0.03	0.04	0.28
PWT8	0.73	0.36	0.01	0.02	0.30
WDI	0.48	0.38	0.01	0.02	0.36
Gabon					
PWT6	0.46	0.09	0.00	0.21	0.55
PWT7	0.07	0.31	0.00	0.05	0.68
PWT8	0.23	0.18	0.00	0.04	0.64
WDI	0.11	0.30	0.00	0.03	0.65
Kenya					
PWT6	0.51	0.28	0.65	0.29	0.52
PWT7	0.45	0.25	0.54	0.50	0.19
PWT8	0.91	0.01	0.36	0.01	0.59
WDI	0.90	0.31	0.17	0.08	0.56
Benin					
PWT6	0.29	0.21	0.01	0.02	0.65
PWT7	0.32	0.23	0.00	0.01	0.58
PWT8	0.07	0.14	0.01	0.01	0.58
WDI	0.11	0.42	0.02	0.02	0.35
Lesotho					
PWT6	0.10	0.00	0.67	0.03	0.63
PWT7	0.57	0.01	0.53	0.66	0.62
PWT8	0.21	0.06	0.02	0.05	0.67
WDI	0.14	0.00	0.05	0.03	0.57
Mauritania					
PWT6	0.10	0.37	0.23	0.06	0.38
PWT7	0.99	0.21	0.63	0.02	0.37
PWT8	0.91	0.41	0.26	0.03	0.38
WDI	0.62	0.18	0.45	0.30	0.55
Togo					
PWT6	0.34	0.55	0.06	0.05	0.44
PWT7	1.00	0.92	0.21	0.17	0.53
PWT8	0.99	0.91	0.16	0.17	0.53
WDI	0.42	0.06	0.24	0.06	0.48

Notes: All reported values are p-values, with the exception of *Trace Corr.*, which is the trace correlation. *Autocorr.*, *Norm.*, and ARCH are tests for no residual autocorrelation, multivariate normality and no ARCH effects respectively, further discussed in the main text.

Source: Author's calculations.

Table 4: Misspecification tests

identical to the second decimal place (meaning that the impact of differences in the aid data is negligible, and the original specification almost certainly relied on the same test statistic). In these cases, no sensible variations in the specification could rectify the issue. This illustrates the limits of heterogeneity even in this comparatively highly flexible framework: A VAR including the present set of variables has limited validity in these cases, and more fundamental changes in the framework may be required to represent the macroeconomic dynamics described by the data. As this concerns both the original study and the re-specifications, it is however of no major concern for the assessment of the robustness of the results.

Similarly, in many models there is some evidence for the presence of ARCH effects (column 3 and 4), both in the PWT6 models consistent with JMT, and in the re-specified ones. This could be another factor undermining the performance of the trace test, which has however been shown to be relatively robust in this respect (Rahbek et al., 2002).

The trace correlation, reported in the last column, is quite persistent within each country and across datasets, indicating that the re-specified models typically reach a similar fit as the ones derived and employed by JMT.

5.2.4. Cointegration rank r

Perhaps the most influential, and often contentious choice, is that of the cointegration rank, that is, the rank r of the long-run coefficient matrix $\mathbf{\Pi}$. Given the short sample period, the standard procedure for the determinations, the trace text (Johansen, 1988) has very low power in the current set-up (JMT, p. 14). It therefore fails to reject unit roots at ranks that are both economically and statistically implausible, meaning that it will tend to indicate unreasonably low ranks.

In line with JMT, a number of criteria are employed in order to assure a well-grounded choice of r. Apart from the aforementioned (i) trace test, I will base my choice of rank on (ii) the largest unrestricted roots of the companion matrix, (iii) the t-ratios of the

α-coefficients, and (iv) a visual inspection of the graphs of the cointegrating relations. These are the same criteria employed by JMT (p. 14); for a more comprehensive discussion, refer to Juselius (2006, chapter 8.5).

As reported in the first column of table 5, and in line with the above mentioned low power of the trace test for cointegration (i), it systematically suggests ranks that are equal to, or lower, than the preferred rank that I eventually determine after consideration of criteria (ii)-(iv). The only exceptions to this occur for Lesotho PWT6, 7 and 8. This is not surprising in the light of the results discussed in the previous section, where the residuals in these models have been found likely to violate the assumption of normality, the fundamental assumption on which the test relies[10]. I emphasise once more that this is pervasive throughout both the re-specified models, as well as those employed by JMT[11].

The (ii) largest unrestricted roots are reported in columns 4 and 5 of table 5, for both the eventually preferred rank r^*, and for $r^* + 1$. The idea here is that the largest unrestricted root in the system should be significantly smaller than one, in order to ensure a stationary process $\Delta\mathbf{X}_t$. Practically speaking, the largest unrestricted root at r^* should be small, while the one at $r^* + 1$ should be close to one, as otherwise the rank could be confidently increased, preserving the information concerning another equilibrium relationship. It is, however, only indicative again, especially because the confidence intervals of the roots are unknown, and there is therefore no hard criterion in order to determine whether a root is significantly smaller than one (Juselius, 2006, p. 143). Nevertheless, the largest roots at $r^* + 1$ are in the vast majority of models substantially larger than those at r^*, providing some justification for this choice. Most importantly in the context of a robustness check, they tend to be of a similar amplitude

[10]The table does not report a p-value in the cases where the trace test concludes full rank $r = p = 5$, as this inference emerges from the rejection of all lower ranks versus the alternative hypothesis of $r = p$.

[11]The relevant output underlying the original specifications, kindly provided by the authors, confirms this.

	Trace test		Largest root		max $\|t\|$ in $\boldsymbol{\alpha}_{r^*}$			
Country / Data	*Inf.*	*p-value*	r^*	r^*+1	r^*	r^*+1	*Graph*	$r^*(r')$
Burkina Faso								
PWT6	0	0.52	0.32	0.86	2.98	3.12	1(2)	2(1)
PWT7	2	0.21	0.59	0.63	-4.1	-4.54	2(1)	2(1)
PWT8	2	0.17	0.54	0.61	-4.17	4.42	2(1)	2(1)
WDI	1	0.18	0.34	0.63	4.51	3.21	2(3)	3(2)
Cameroon								
PWT6	0	0.08	0.63	0.74	-4.66	1.34	3(2)	3(2)
PWT7	0	0.55	0.64	0.68	-2.44	-1.18	2(3)	3(2)
PWT8	0	0.25	0.66	0.69	2.55	1.44	2(3)	3(2)
WDI	1	0.11	0.6	0.73	3.84	-2.59	3(2)	3(2)
Gabon								
PWT6	1	0.21	0.92	0.92	4.16	-2.89	3(4)	3(4)
PWT7	3	0.19	0.38	0.44	7.66	-3.46	3(4)	3(4)
PWT8	2	0.16	0.87	0.99	-3.67	-4.57	3(4)	3(2)
WDI	2	0.08	0.93	0.99	5.92	-2.98	3(4)	3(2)
Kenya								
PWT6	1	0.2	0.61	0.83	4.63	4.31	2(3)	3(2)
PWT7	1	0.06	0.67	0.85	-2.86	-3.71	2(3)	3(2)
PWT8	3	0.22	0.55	0.85	3.53	-4.49	2(3)	3(2)
WDI	3	0.08	0.64	0.78	4.41	5.02	4(3)	3(4)
Benin								
PWT6	3	0.23	0.36	0.65	-4.83	3.18	3(4)	3(4)
PWT7	2	0.33	0.31	0.82	4.23	-1.45	3(2)	3(2)
PWT8	2	0.1	0.24	0.63	3.97	-3.12	3(2)	3(2)
WDI	0	0.16	0.25	0.35	-4.34	-3.61	2(1)	2(1)
Lesotho								
PWT6	5	-	0.85	0.89	4.6	4.46	1(2)	3(2)
PWT7	4	0.07	0.55	0.83	5.02	-3.53	2(1)	2(1)
PWT8	5	-	0.55	0.6	5.76	5.01	2(3)	3(2)
WDI	2	0.37	0.58	0.82	5.34	-3.15	1(2)	2(1)
Mauritania								
PWT6	0	0.23	0.53	0.63	-3.17	-2.68	2(3)	3(2)
PWT7	0	0.2	0.41	0.77	-4.54	-2.94	3(2)	3(2)
PWT8	0	0.33	0.45	0.62	-4.27	-4.08	2(3)	2(3)
WDI	2	0.41	0	0.61	4.09	3.28	2(1)	2(1)
Togo								
PWT6	2	0.41	0.34	0.68	-4.96	-3.01	2(1)	2(3)
PWT7	3	0.26	0.59	0.66	4.53	-1.85	3(2)	3(2)
PWT8	3	0.17	0.63	0.68	-4.93	1.95	3(2)	3(2)
WDI	2	0.62	0.54	0.78	3.43	1.62	3(2)	3(2)

Notes: *Trace test* reports the rank suggested by Johansen (1988)'s trace test with the corresponding p-value of acceptance. *Largest root* and *max* $|t|$ *in* $\boldsymbol{\alpha}_{r^*}$ report the respective values for the preferred rank and the one above, *Graph* indicates the rank most confidently suggested by the graph and the best alternative in parentheses, $r^*(r')$ the inferred preferred and second best choice of rank.
Source: Author's calculations.

Table 5: Criteria for the choice of rank

as those obtained in the PWT6 (JMT) models.

The (iii) largest t-ratio in the α-vector associated with the r^*'th CI relation (column 5) gives an indication about the relevance of the last potential equilibrium relationship included in the model. The same figure for the r^*+1'th α-vector (column 6) provides such an indication for the first vector dismissed from the analysis. Juselius (2006) proposes a threshold of about $|t| > 2.6$, which is surpassed in most specifications, as $r^* + 1$ is typically found to be susceptible of non-stationarity. Where this is not the case, $r^* + 1$ is generally determined as the second best choice of rank, reported in parentheses in the last column.

In practice, the (iv) visual inspection of the graphs of the CI relationships turn out to be the most common tie-breaker. While in principle, this is quite a subjective criterion, it turns out that in most cases there tends to be a rather sharp difference in the appearance of the graphs of the r^*'th CI (the CIs being ordered by the corresponding eigenvalues) and the r^*+1'th (or, if this is the second best choice, the r^*+2'th). The highest included CI typically looks quite a lot like a white noise process, while larger order CIs have a distinctively persistent, that is, non-stationary appearance. The values reported in the second last column of table 5 reports the rank that can be best justified based on the graphs, followed by the best alternative rank choice resulting from them in parentheses; In cases where the difference is quite clear-cut, the alternative rank will be below the preferred choice. Where the next-highest CI also appears to be acceptable, e.g., has a short period of persistence but otherwise looks stationary, it may be reported as the suggested second best rank.

The last column reports the final choice of ranks after weighting of criteria (i)-(vi)[12].

[12] The choice of 3(2) for Togo PWT6, a model repeatedly found to be problematic in the previous sections, may deserve some discussion as it seems at odds with the indications provided by the criteria. The main rationale behind the choice of rank is to preserve consistency with JMT, even though their choice may appear rather surprising in this particular instance. As noted earlier, JMT's output is very much in line with the results I obtain; this can be verified for the trace test and the roots of the companion matrix.

It is apparent from the discussion in this section that the choice of the cointegration rank is everything but straightforward, and the researcher faces significant trade-offs in the process. This provides justification for JMT's procedure of assessing the results from two different economic angles, establishing transparency by essentially picking the rank that yields the results most consistent with the respective prior of effectiveness or harmfulness, and reporting both.

5.2.5. Results of the re-specified models

Table 6 reports the final model choices, stating their respective lag-length, dummy variables, and first and second best choices of cointegration ranks.

The results emerging from these specification are summarised in table 7, the underlying t-ratios are reported in appendix B. For comparability, the notation has been adopted from JMT (table 5), and some key figures have been included in order to facilitate the assessment of the relative stability / instability of the results within the countries and across the datasets. As in JMT, a $+$ corresponds to a positive t-ratio larger than 2, which is considered a case of aid effectiveness, while $-$ indicates a t-ratio smaller than -2, and thus harmfulness. The subscript 0 indicates absolute t-ratios between 2 and 1.6 (marginally (in)significant), the subscript 00 an absolute t-ratio smaller than 1.6 (insignificance). Under the prior of aid effectiveness, aid is considered to have been *effective* in a country if the coefficient of either Y or I is positive significant ($+$). Conversely, under the prior of harmfulness, a negative significant coefficient ($-$) on either of the two leads to the conclusion of overall harmfulness. The table is organised in two columns for each dataset, reporting the results under the prior of effectiveness and the prior of harmfulness respectively, and four rows per country, one for each of the macrovariables under consideration.

Country	*Lags*	*Dummy variables*	r^*	r'
		Stable countries		
Burkina Faso		*Sample: 1965-2007*		
PWT6	1	$Dtr71_t, Dp00_t$	2	1
PWT7	1	$Dtr71_t, Ds74_t, Dp79_t, Dp85_t, Dp94_t, Dp00_t$	2	1
PWT8	1	$Dtr71_t, Ds74_t, Dp79_t, Dp85_t, Dp94_t, Dp00_t$	2	1
WDI	1	$Dp85_t, Dp94_t, Dp00_t$	2	3
Cameroon		*Sample: 1965-2007*		
PWT6	1	None	3	2
PWT7	1	None	3	2
PWT8	1	None	3	2
WDI	1	None	3	2
Gabon		*Sample: 1965-2002*		
PWT6	1	$Dtr7478_t Dtr8900_t$	3	4
PWT7	1	$Dtr7478_t Dp87_t Dp00_t$	3	2
PWT8	1	$Dtr7478_t Dp87_t Dtr0002_t$	3	2
WDI	1	$Dtr7478_t Dp87_t Dtr0002_t$	3	2
Kenya		*Sample: 1965-2007*		
PWT6	1	$Ds68_t, Ds94_t$	3	2
PWT7	1	$Ds68_t, Ds94_t$	3	2
PWT8	1	$Ds68_t, Dtr9900_t, Ds94_t$	3	2
WDI	1	$Ds70_t, Ds94_t$	3	4
		Unstable countries		
Benin		*Sample: 1965-2007*		
PWT6	1	$Dp75_t, Ds82_t, Dp94_t$	3	4
PWT7	1	$Dp82, Ds86, Dp94$	3	2
PWT8	1	$Ds82_t, Dp89_t, Dp94_t$	3	2
WDI	1	None	2	1
Lesotho		*Sample: 1963-2007*		
PWT6	1	$Dtr65t, Ds70t, Ds78_t, Dp83t, Dp99t$	3	2
PWT7	1	$Ds71_t, Dp78_t, Dtr65_t, Dp83_t, Dp99_t$	2	1
PWT8	1	$Ds71_t, Dp75_t, Ds78_t, Dp83_t, Dp99$	3	2
WDI	1	$Ds71_t, Dp75_t, Ds78_t, Dp83_t, Dp98, Dp99$	2	1
Mauritania		*Sample: 1965-2007*		
PWT6	1	$Ds92_t$	3	2
PWT7	1	$Dp86_t$	3	2
PWT8	1	$Dp86_t, Dp95_t$	2	3
WDI	1	$Dp76, Dp99_t, Dp03_t$	2	1
Togo		*Sample: 1965-2007*		
PWT6	1	$Dp93_t$	2	3
PWT7	1	$Dp86_t, Dp87_t, Dp93_t$	3	2
PWT8	1	$Dp86_t, Dp87_t, Dp93_t$	3	2
WDI	1	$Dp71_t, Dp75_t, Dp93_t$	3	2

Table 6: Final model specifications

		PWT6		PWT7		PWT8		WDI		Stable	
		Eff.	Harm.	Eff.	Harm.	Eff.	Harm.	Eff.	Harm.	Eff.	Harm.
Burkina Faso	Y	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$+_{00}$	3	2
	I	$+_{00}$	$+_{00}$	$+_{00}$	$-$	$+_{00}$	$+_{00}$	$-_{00}$	$-$	2	1
	C	$-_{00}$	$-_{00}$	$-_{00}$	$+_{0}$	$-_{00}$	$-_{00}$	$+_{00}$	$+$	3	1
	G	$+_{00}$	$+_{00}$	$-_{00}$	$+_{0}$	$+_{00}$	$+_{00}$	$+_{0}$	$-$	2	1
Cameroon	Y	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$-$	3	2
	I	$+_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$+_{00}$	$+_{00}$	3	3
	C	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$-$	3	2
	G	$-_{00}$	$-$	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	3	0
Gabon	Y	$+_{00}$	$+_{00}$	$+_{00}$	$-_{00}$	$+_{0}$	$+_{0}$	$+_{00}$	$-_{00}$	2	2
	I	$+_{00}$	$-_{00}$	$+_{00}$	$-_{00}$	$+_{0}$	$+_{0}$	$+_{00}$	$-_{00}$	2	2
	C	$+_{00}$	$-$	$-_{00}$	$-_{00}$	$+_{00}$	$+_{00}$	$-_{00}$	$-_{00}$	3	0
	G	$+_{00}$	$+_{00}$	$+_{00}$	$-_{00}$	$+_{00}$	$+_{00}$	$+_{00}$	$-_{00}$	3	3
Kenya	Y	$+$	$+_{00}$	$+$	$+$	$+$	$+$	$+$	$+$	3	0
	I	$+$	$+$	$+$	$+_{0}$	$+$	$+_{0}$	$+$	$+$	3	1
	C	$+$	$+_{0}$	$+$	$+$	$+$	$+$	$+$	$+$	3	0
	G	$+_{00}$	$+_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	$+_{0}$	$-$	2	2
Benin	Y	$+_{00}$	$-$	$+$	$-_{00}$	$+$	$-_{00}$	$-_{00}$	$-_{0}$	1	0
	I	$+$	$-$	$+$	$-$	$+$	$+_{00}$	$+_{00}$	$-$	2	2
	C	$+$	$+$	$+$	$+$	$+$	$+$	$-_{00}$	$+_{00}$	2	2
	G	$-$	$-$	$+$	$-_{00}$	$-$	$-$	$+_{00}$	$+$	1	1
Lesotho	Y	$+$	$+$	$-$	$-$	$+_{00}$	$-_{00}$	$+_{00}$	$-$	0	0
	I	$+$	$+$	$+_{00}$	$+_{00}$	$+$	$+_{00}$	$+_{00}$	$-$	1	0
	C	$+$	$+$	$-$	$-$	$-$	$-$	$+_{00}$	$-$	0	0
	G	$+$	$+_{00}$	$-_{00}$	$-_{00}$	$+_{00}$	$-$	$-_{0}$	$-_{00}$	0	2
Mauritania	Y	$-_{00}$	$+_{00}$	$+_{0}$	$-$	$-_{00}$	$-$	$-$	$-$	1	0
	I	$+$	$-_{00}$	$+_{00}$	$-$	$+_{00}$	$-_{00}$	$-_{00}$	$-_{00}$	0	2
	C	$+_{00}$	$+_{00}$	$+$	$-_{00}$	$+$	$+$	$+_{00}$	$+_{00}$	1	2
	G	$+$	$+_{00}$	$+_{00}$	$+$	$+_{00}$	$+_{00}$	$+$	$+$	1	1
Togo	Y	$+$	$+_{0}$	$+$	$+$	$+$	$+$	$+$	$+$	3	0
	I	$-_{00}$	$+_{0}$	$-$	$-$	$-$	$-$	$+$	$+_{00}$	0	0
	C	$+$	$-_{00}$	$+$	$+$	$+$	$+$	$+$	$+$	3	0
	G	$+$	$+$	$+_{00}$	$+_{00}$	$+_{00}$	$+_{00}$	$+_{00}$	$+_{00}$	0	0
	Y	3	1	3	2	3	1	2	3		
Σ	I	4	1	2	4	3	1	2	3		
	Σ	5	1	3	5	4	2	2	5		

Notes: + indicates a positive coefficient, − a negative one. The subscript '0' indicates an absolute t-ratio between 1.6 and 2, '00' indicates one lower than 1.6. The absence of a subscript indicates an absolute t-ratio > 2. The last three rows count the number of cases of effectiveness and harmfulness on Y, I, and overall. The last two columns count the cases of consistency under each prior.

Table 7: Results of the re-specified models

The last two columns of table 7 count the number of coefficients that are consistent with the ones obtained from PWT6[13], consistency again meaning that they are identical with regards to sign and the level of significance (i.e., the t-ratio falls within the same interval), or are insignificant ($|t| < 1.6$) in either instance, regardless the sign. An apparent pattern is that the countries yielded particularly consistent results under the original JMT models also remain substantially more consistent under the re-specified models, with 43 out of 48 consistent coefficients under the prior of aid effectiveness, compared to 16 in the countries with the least consistent results in the previous exercise.

This pattern is repeated under the prior of harmfulness, but slightly less pronounced and generally at a lower level of consistency. The datasets agree for 22 coefficients in the stable countries, and for 12 in the least stable ones. Note that in the first exercise, the results for this sub-sample of countries were approximately equally consistent, with a total of 62 out of 96 coefficients consistent under the prior of effectiveness, and 60 under that of harmfulness. Also, the distribution of consistent coefficients across countries was roughly the same, with 48 of them being in the consistent countries under the prior of effectiveness, and 44 under the prior of harmfulness.

One caveat is that the results seem to be systematically different between stable and unstable countries, in the sense that most of the coefficients of the stable countries are in fact insignificant. Note however that even Kenya, where 3 out of 4 coefficients are significant with PWT6, is almost perfectly consistent under the prior of effectiveness, and far better than any of the inconsistent countries.

The conclusions emerging from these results are summarised in the last three rows of table 7 for each of the four datasets, across the 8 countries in the sub-sample. The first two of these rows report the number of positive (negative) significant coefficients

[13]Note that while these are overall consistent with the results obtained by JMT, but may differ in some cases; where this is the case, I use my own results as a benchmark for the sake of internal consistency. This affects mainly Burkina Faso, where differences in the aid data also lead me to a different specification than JMT, and the coefficients of investment and consumption in Togo.

for GDP (Y) and investment (I) respectively. The last row (Σ) counts the number of countries where at least one of the two is significant and positive (negative), and where therefore aid is considered to have been effective (harmful) in the long run, overall. This reveals that, in the sub-sample at hand, the overall conclusions change twice once we account for the impact of the data on the modelling process, and remain constant only once, but with less strong support. The conclusion of aid effectiveness, clearly supported by PWT6 with 5 countries providing evidence for it, compared to only 1 country (Benin) providing evidence for harmfulness, only finds support in PWT8, where 4 countries indicate effectiveness, and 2 harmfulness. PWT7 and WDI now lend some support to the hypothesis of harmfulness, with 5 countries providing evidence for it in each dataset, compared to 3 (PWT7) and 2 (WDI) cases of effectiveness.

6. Conclusions

In an attempt to assess the stability of macroeconomic inference drawn from methods of Time Series Analysis, especially in the context o foreign aid effectiveness, I have used Juselius, Møller and Tarp (2014) as a framework for extensive robustness chess with respect to the data.

This included, in a first exercise, the application of the models as specified by JMT to three alternative but conceptually similar datasets. About one third of the coefficients changed qualitatively in each of the datasets applied, the country-wise inference (aid effectiveness, harmfulness or neutrality) similarly remained stable in about two thirds of the country-dataset combinations. It is worth noting that there is some pattern of clustering between the countries, with almost half of our sample (17 countries) yielding consistent inference throughout the replications, concentrating all the inconsistencies on the remaining half of the sample.

The second exercise allowed for a more fundamental impact of the data as I re-specified the country-specific models for each datasets for the four least and the four most stable countries respectively. The results show that the modelling process can be influenced in a significant manner by the inconsistencies between the datasets, especially when it comes to the detection of outliers and the resulting dummy variables, and to the choice of the rank of cointegration. The results follow a predictable pattern: the countries that proved to be particularly consistent in the first exercise also yielded less divergent models in the second exercise, and consequently more similar results. In the countries that were particularly inconsistent in the first exercise however, the alterations to the models had as a main result that the significance of the results was restored (likely due to reducing the noise by accounting for the appropriate outliers). These results do not necessarily correspond to the original results obtained by JMT, and in many of our cases, the re-specification therefore exacerbated the divergence.

One obvious conclusion to draw from this exercise is that the choice of data matters, and robustness checks with alternative data should become standard in the literature, wherever feasible. This is not always straightforward to perform, especially in the context of Time Series Analysis, as the specification may have to be rethought when the data changes; however, this only increases the potential ramifications of changes in the data and renders the exercise even more urgent.

The issue of data quality and measurement error occurs at different levels. I have not touched here the basic level of the validity of the original measurements, but instead focussed on the impact of second-order modifications by data suppliers. The impact these modifications can have, illustrated by my results, underlines the necessity of tractable methodological choices and documentation on part of data suppliers.

However, it needs to be highlighted that for about half of the countries included in the analysis, the results were largely unaffected by the modifications in the data. Categorically dismissing the only available evidence on their past economic performances, as advocated by some, would represent a considerable waste of information; it is what the best estimates tell us, and the reported sums for these countries are sufficiently in agreement in order to provide credible insights about the mechanisms at work.

References

Boone, P. (1996). Politics and the effectiveness of foreign aid. *European Economic Review*, 40(2):289–329.

Breton, T. R. (2012). Penn world table 7.0: Are the data flawed? *Economics Letters*, 117(1):208–210.

Brückner, M. (2013). On the simultaneity problem in the aid and growth debate. *Journal of Applied Econometrics*, 28(1):126–150.

Burnside, C. and Dollar, D. (2000). Aid, policies, and growth. *The American Economic Review*, pages 847–868.

Dalgaard, C.-J., Hansen, H., and Tarp, F. (2004). On the empirics of foreign aid and growth*. *The Economic Journal*, 114(496):F191–F216.

Dollar, D. and Easterly, W. (1999). The search for the key: aid, investment and policies in Africa. *Journal of African Economies*, 8(4):546–577.

Doornik, J. A. (2009). Autometrics. In Castle, J. and Shepard, N., editors, *The Methodology and Practice of Econometrics*, chapter 4, pages 88–121. Oxford University Press.

Easterly, W. (2003). Can foreign aid buy growth? *The Journal of Economic Perspectives*, 17(3):23–48.

Easterly, W., Levine, R., and Roodman, D. (2004). Aid, policies, and growth: Comment. *The American Economic Review*, 94(3):774–780.

Engle, R. F. and Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica*, pages 251–276.

Feenstra, R. C., Inklaar, R., and Timmer, M. (2013). The next generation of the Penn World Table. *University of California, Davis and University of Groningen.*

Gebregziabher, F. (2014). The long-run macroeconomic effects of aid and disaggregated aid in Ethiopia. *Journal of International Development*, 26(4):520–540.

Guha-Sapir, D., Below, R., and Hoyois, P. (2014). EM-DAT: International disaster database. *Univ. Cathol. Louvain, Brussels: Belgium.*

Hansen, H. and Doornik, J. A. (1994). An omnibus test for univariate and multivariate normality. *Economics Papers Series (Nuffield College, University of Oxford).*

Hansen, H. and Tarp, F. (2001). Aid and growth regressions. *Journal of Development Economics*, 64(2):547–570.

Herzer, D. and Morrissey, O. (2013). Foreign aid and domestic output in the long run. *Review of World Economics*, 149(4):723–748.

Heston, A., Summers, R., and Aten, B. (2009). Penn World Table v. 6.3. *Center for International Comparisons of Production, Income and Prices (Philadelphia: University of Pennsylvania).*

Heston, A., Summers, R., and Aten, B. (2012). Penn World Table version 7.1. *Center for International Comparisons of Production, Income, and Prices at the University of Pennsylvania.*

Hoover, K. D., Johansen, S., and Juselius, K. (2008). Allowing the data to speak freely: The macroeconometrics of the cointegrated vector autoregression. *The American Economic Review*, 98(2):251–255.

Hsieh, C.-T. and Klenow, P. J. (2003). Relative prices and relative prosperity. Technical report, National Bureau of Economic Research.

Jerven, M. (2011). Users and producers of African income: Measuring the progress of African economies. *African affairs*, 110(439):169–190.

Jerven, M. (2013a). Briefing: For richer, for poorer: GDP revisions and Africa's statistical tragedy. *African Affairs*, 112(446):138–147.

Jerven, M. (2013b). Comparability of GDP estimates in sub-Saharan Africa: The effect of revisions in sources and methods since structural adjustment. *Review of Income and Wealth*.

Jerven, M. (2013c). *Poor Numbers: How We Are Misled by African Development Statistics and What to Do about It*. Cornell University Press.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, 12(2):231–254.

Johansen, S. (1996). *Likelihood Based Inference in Cointegrated Vector Autoregressive Models*. Advanced Texts in Econometrics. Oxford University Press.

Johnson, S., Larson, W., Papageorgiou, C., and Subramanian, A. (2012). Is newer better? Penn World Table revisions and their impact on growth estimates. *Journal of Monetary Economics*.

Juselius, K. (2006). *The cointegrated VAR model: methodology and applications*. Oxford University Press.

Juselius, K., Møller, N. F., and Tarp, F. (2014a). 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.

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.*

Kobou, G., Njinkeu, D., and Powo Fosso, B. (2008). The political economy of Cameroon's post-independence growth experience. In Ndulu, B. J., O'Connell, S. A., Bates, R. H., Collier, P., and Soludo, C. C., editors, *The Political Economy of Economic Growth in Africa, 1960–2000*, chapter 16, pages 547–587. Cambridge University Press.

Martins, P. M. (2010). Fiscal dynamics in Ethiopia: The cointegrated VAR model with quarterly data. Technical report, CREDIT Research Paper.

Morrissey, O. (2001). Does aid increase growth? *Progress in Development Studies*, 1(1):37–50.

Morrissey, O., M'Amanja, D., and Lloyd, T. (2007). Aid and growth in Kenya: A time series approach. In Lahiri, S., editor, *Theory and practice of foreign aid*, volume 1 of Frontiers of economics and globalization, pages 313–332. Elsevier.

Ndulu, B. J., O'Connell, S. A., Bates, R. H., Collier, P., and Soludo, C. C., editors (2008). *The political economy of economic growth in Africa, 1960-2000*, volume 2. Cambridge University Press.

Osei, R., Morrissey, O., and Lloyd, T. (2005). The fiscal effects of aid in Ghana. *Journal of International Development*, 17(8):1037–1053.

Pettersson, T. and Wallensteen, P. (2015). Armed conflicts, 1946–2014. *Journal of Peace Research*, 52(4):536–550.

Ponomareva, N. and Katayama, H. (2010). Does the version of the Penn World Tables matter? An analysis of the relationship between growth and volatility. *Canadian Journal of Economics/Revue canadienne d'économique*, 43(1):152–179.

Rahbek, A., Hansen, E., and Dennis, J. G. (2002). ARCH innovations and their impact on cointegration rank testing. *Centre for Analytical Finance working paper*, 22:15.

Rajan, R. G. and Subramanian, A. (2008). Aid and growth: What does the cross-country evidence really show? *The Review of Economics and Statistics*, 90(4):643–665.

Rajan, R. G. and Subramanian, A. (2011). Aid, dutch disease, and manufacturing growth. *Journal of Development Economics*, 94(1):106–118.

Ram, R. and Ural, S. (2013). Comparison of GDP per capita data in Penn World Table and World Development Indicators. *Social Indicators Research*, pages 1–8.

Ramey, G. and Ramey, V. (1995). Cross-country evidence on the link between volatility and growth. *The American Economic Review*, 85(5):1138–1151.

Roodman, D. (2007). The anarchy of numbers: aid, development, and cross-country empirics. *The World Bank Economic Review*, 21(2):255–277.

Seely, J. C. (2005). The legacies of transition governments: post-transition dynamics in Benin and Togo. *Democratization*, 12(3):357–377.

Shenton, L. and Bowman, K. (1977). A bivariate model for the distribution of $\sqrt{b1}$ and b2. *Journal of the American Statistical Association*, 72(357):206–211.

Temple, J. R. (2010). Chapter 67 - aid and conditionality*. In Rodrik, D. and Rosenzweig, M., editors, *Handbooks in Economics*, volume 5 of *Handbook of Development Economics*, pages 4415 – 4523. Elsevier.

Temple, J. R. and Van de Sijpe, N. (2014). Foreign aid and domestic absorption.

The World Bank (2015). World Development Indicators.

APPENDIX

A. Consistency by country

Country	*PWT6*	*PWT7*	*PWT8*	*WDI*	*Consistent*
BDI	4	4	2	-	10/12
BEN	**4**	**2**	**2**	**2**	**10/16**
BFA	**4**	**4**	**4**	**4**	**16/16**
BWA	1	0	0	-	1/12
CAF	4	4	0	-	8/12
CMR	**4**	**4**	**4**	**4**	**16/16**
COG	4	4	3	3	14/16
COM	4	0	3	-	7/12
DJI	2	2	0	-	4/12
ETH	4	1	0	-	5/12
GAB	**4**	**4**	**4**	**4**	**16/16**
GHA	4	2	1	-	7/12
GIN	4	3	4	-	11/12
GMB	4	1	3	-	8/12
KEN	**4**	**4**	**4**	**4**	**16/16**
LBR	4	4	0	-	8/12
LSO	**4**	**0**	**0**	**0**	**4/16**
MDG	4	4	3	3	14/16
MLI	3	3	3	-	9/12
MRT	**1**	**0**	**1**	**2**	**4/16**
MUS	4	3	3	-	10/12
MWI	4	3	3	-	10/12
NER	4	3	4	-	11/12
NGA	4	2	2	-	8/12
RWA	3	2	3	2	10/16
SDN	3	2	2	3	10/16
SEN	3	4	4	1	12/16
SOM	4	3	0	-	7/12
SWZ	4	2	2	-	8/12
SYC	3	4	0	-	7/12
TCD	4	4	4	-	12/12
TGO	**2**	**2**	**2**	**1**	**7/16**
TZA	3	1	1	-	5/12
UGA	4	1	2	-	7/12
ZMB	3	2	2	-	7/12
ZWE	4	1	1	-	6/12

The table reports the number of coefficients that are consistent with those obtained by JMT under the prior of aid effectiveness, taking into consideration the first an second best choice of rank, using their exact models. The last column reports the total sum of consistent coefficients within each country across all datasets, followed by the number of estimated coefficients. Countries included in the re-specification exercise are in bold.

Table 8: Consistent coefficients by country

B. T-ratios of the C-Matrix

	PWT6		PWT7		PWT8		WDI	
	r^*	r'	r^*	r'	r^*	r'	r^*	r'
Burkina Faso	-1.03	-1.08	-1.08	-0.35	-1.03	-1.08	0.65	-0.24
	1.15	-0.13	0.68	-4.00	1.15	-0.13	-3.81	-1.52
	-0.26	-0.43	-0.29	1.78	-0.26	-0.43	2.36	0.72
	0.16	0.89	-0.41	1.70	0.16	0.89	-2.37	1.68
Cameroon	-0.55	-1.19	-0.91	-0.93	-0.89	-0.68	-1.24	-3.01
	0.62	-0.30	-0.22	-1.12	-0.21	-1.20	0.09	0.12
	-0.15	-0.59	-0.82	-0.26	-0.80	-0.03	-0.28	-2.07
	-0.45	-3.85	-0.61	-0.54	-0.27	0.73	-1.48	-0.78
Gabon	0.23	1.02	-0.41	0.14	1.89	1.81	-0.50	0.65
	-0.01	1.02	-0.41	0.19	1.88	1.81	-0.59	0.57
	-2.80	1.02	-0.42	-0.02	0.68	1.22	-0.88	-0.69
	1.18	1.02	-0.42	0.12	1.32	1.55	-0.86	0.77
Kenya	4.03	1.44	2.54	2.84	2.69	3.02	4.75	2.23
	3.38	3.09	2.01	1.88	2.46	1.82	4.53	2.23
	3.83	1.80	2.59	2.96	2.72	3.04	4.41	2.23
	0.67	1.05	-1.13	-0.54	-0.18	-0.15	1.69	-2.23
Benin	1.15	-2.37	-0.63	6.70	-0.89	2.31	-0.75	-1.81
	3.05	-2.37	-4.26	12.78	0.15	9.97	0.70	-4.15
	3.74	2.37	4.19	4.58	2.48	2.43	-1.19	0.76
	-2.84	-2.37	-1.03	6.43	-2.10	-4.38	0.90	2.54
Lesotho	2.34	5.11	-5.94	-11.51	-0.12	0.83	1.09	-7.04
	2.22	2.21	1.56	-1.57	0.93	3.72	1.10	-2.94
	3.28	3.46	-6.90	-6.15	-2.55	-7.35	1.47	-2.50
	2.27	0.02	-1.08	-2.43	-7.02	0.49	-1.95	-1.00
Mauritania	-1.30	0.00	1.85	-3.64	-0.93	-2.97	-5.20	-0.31
	2.17	-0.26	0.18	-2.91	0.75	-0.02	-0.75	-26.12
	0.58	0.06	4.51	-0.51	2.93	4.80	0.27	0.91
	2.84	0.08	0.97	3.06	0.90	0.41	2.38	-1.78
Togo	3.64	1.99	11.16	2.47	11.63	1.49	4.22	3.66
	-0.94	1.73	-3.62	-4.98	-3.55	-4.81	2.29	0.81
	4.88	-1.41	5.83	1.63	4.26	1.04	6.45	6.61
	2.22	3.73	1.11	1.13	0.37	0.58	0.88	0.68

Notes: r^* reports the t-ratios for the variables in the second column under the preferred rank specification, r' those obtained under the second best choice of rank.
Source: Author's calculations.

Table 9: t-ratios of the best and second best choice of rank