Government Revenue Dataset (GRD)

Government Revenue Dataset (2021): source selection

Kyle McNabb,[1] Annalena Oppel,[2] and Daniel Chachu[3]

August 2021

wider.unu.edu

Abstract: This technical note is part of a series of technical notes describing the construction of the Government Revenue Dataset (GRD). We provide an overview of the underlying sources in the GRD and the criteria used in order to select data from them. In addition, details of the approach taken to merge sources within countries and across government levels are given. We also discuss exceptions to the source selection criteria.

Key words: GRD, source selection, government revenue, tax revenue data

JEL classification: C82, H20, E00

Related publications:

Oppel, A., K. McNabb, and D. Chachu (2021). 'Government Revenue Dataset (2021): Variable Description'. WIDER Technical Note 2021/11. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-11

McNabb, K., A. Oppel, and D. Chachu (2021). 'Government Revenue Dataset (2021): Country Notes'. WIDER Technical Note 2021/[forthcoming]. Helsinki: UNU-WIDER.

Data: Government Revenue Dataset

[1] ODI, London, UK, corresponding author k.mcnabb@odi.org.uk; [2] UNU-WIDER, oppel@wider.unu.edu; [3] independent researcher, doc.chachu@gmail.com

Information and requests: publications@wider.unu.edu

https://doi.org/10.35188/UNU-WIDER/WTN/2021-10

1   Introduction

The Government Revenue Dataset (GRD) (UNU-WIDER 2021) aims to encourage the use of the data in analysis and research to answer questions concerning domestic revenue mobilization, especially in the context of developing countries. These analyses and research serve to inform policies to assist developing countries to raise as much revenue as possible, whilst also helping to stimulate growth and reduce poverty. The dataset currently covers 196 countries, over the period 1980–2019. There are 40 tax and revenue variables in total.

The GRD provides data on government tax and non-tax revenues, social contributions and grants in both local currency and as a percentage of GDP. It also highlights the portion of government revenues that accrue from natural resource extraction. The GRD is a cross-country dataset which complements data from several underlying sources—such as the OECD's Revenue Statistics, and the IMF's Government Finance Statistics—with data from IMF Article IV country reports. This has led to considerable gains in coverage, particularly for developing countries.

An important initial motivation behind the creation of the GRD was a concern about the quality and transparency of data from publicly available data. Public data can exhibit significant limitations, while private data used in internal research can be more informative, yet access is limited. The GRD thus seeks to fill this gap by providing publicly available, transparent, and detailed data to researchers. In doing so, we also highlight data gaps and needs in terms of data quality and access.

For many countries in the GRD, a number of sources (as listed above and further discussed below) are available from which to draw data. This technical note describes the criteria employed in selecting sources for such countries.

2   Data selection criteria in the GRD: an overview

In this section, we describe guiding principles and criteria applied in the selection of revenue data sources. There is no formal data algorithm, or choice ranking, for selecting the source of data for any one country-year observation.[1] Thus, an element of informed subjectivity is necessary in the source choice. However, (i) this is carried out according to a number of criteria, which are explored below and (ii) we describe how this as a distinct strength, rather than a limitation, of the GRD.

The shortcomings of some of the underlying GRD source data are well documented—e.g., Prichard et al. (2014); McNabb (2017)—and therefore necessitate that each and every observation has been examined for quality, consistency, completeness and accuracy. Whilst there are merits to a more formal algorithmic approach or a strict ranking of preference regarding sources, the reasons why this is infeasible for the GRD are laid out in the following. There are a number of criteria that were considered when selecting the source for any one country-year observation. These are summarised in Table 1. Note that the order according to which criteria are listed in the table below do not represent a strict hierarchy (whereby non-fulfilment would lead to subsequent exclusion) but rather show our steps of considerations when selecting sources.

---

[1] Some of the difficulties with this are explored in Prichard et al. (2014).

Table 1: Overview of source selection criteria

Criteria	Brief description
'Length' of coverage[2]	This criterion considers how far the time series goes back. A source with longer coverage is preferred.
'Breadth' of coverage	The level of disaggregation of revenue components. A source reporting more subcomponents is preferred.
Availability of data on natural resource revenues	Where revenue (both tax and non-tax, if available) is reported according to whether it accrues from natural resources, this is preferred.
Availability of data on receipts from mandatory social security contributions.	If a source reports Social Security Contributions as a subcomponent of Total Revenue, then this is preferred.
Level of government reporting	General government is normally preferred to central government, as it captures more revenue. If central has fulfilled more of the other criteria than general and the amount of revenue reported at other (non-central) levels of government is negligible.
Consistency across sources	Do the levels and trends in the source match those from other sources? If not, then the source is unlikely to be included.

Source: authors' compilation.

3   Implementation of selection criteria: data sources

The GRD includes data on government revenues from the following sources:[3]

- OECD *Revenue Statistics*: OECD Member Countries
- OECD *Revenue Statistics:* African Countries
- OECD *Revenue Statistics:* Asian and Pacific Economies
- OECD *Revenue Statistics:* Latin American Countries
- IMF *Government Finance Statistics* (GFS)
- IMF Article IV Staff Reports

Each of these is discussed below, including a brief overview of its relative merits and weaknesses according to the criteria outlined above in Table 1.

3.1   OECD Revenue Statistics

The OECD Revenue Statistics source is one of the primary sources employed in the GRD. Four datasets exist under the *Revenue Statistics* umbrella, and these are discussed in turn.

- OECD *Revenue Statistics*: OECD Member Countries

For OECD member countries, the OECD Revenue Statistics series is often the most complete source in terms of both length and breadth of data coverage. It includes social security

---

[2] For the purposes of the GRD, we are concerned only with data back to 1980.

[3] Data was previously included from the CEPALSTAT *Revenue Statistics in Latin America* dataset, although this now offers few advantages compared to the OECD Revenue Statistics: Latin American Countries dataset, as the two databases are compiled jointly. Data from the OECD comes from stats.oecd.org, the GFS from data.imf.org and the IMF Article IV Staff reports from https://www.imf.org/en/Publications/SPROLLs/Article-iv-staff-reports

contributions and is reported, where available, at both general and central government levels.[4] Over time, the country-level tables have included data on tax revenue received from natural resources. The major shortcoming of the OECD Member Countries Revenue Statistics is the absence of Total Revenue and Non-Tax Revenue data. Only tax and social contributions are reported on. However, the GRD also incorporates data from the OECD National Accounts data series, which allows us for most years to gain a more complete picture of how total government revenue is broken down, although this series often extends only as far back as the mid-1990s.

- OECD *Revenue Statistics:* African Countries

The OECD Revenue Statistics for African Countries is a relatively new series of data and hence was not available at the inception of the GRD. As of 2021, the OECD database has grown to include data for 30 African countries, most often at the central government, but also at general government level where available. The length of coverage varies by country from as far back as 1990 until the mid-2000s. The breadth of data shows a high level of disaggregation where almost all sub-components of revenue, including social contributions, non-tax and grants are reported. The series is now a particularly useful source for resource revenues for African countries

- OECD *Revenue Statistics:* Asian and Pacific Economies

As with the African Countries series, the OECD *Revenue Statistics* for Asian and Pacific Economies has been growing over the past few years, and now includes data for 21 economies in the region. Again, all of the subcomponents of total revenue are reported including resource revenues where applicable. Data is reported at either central, or general government level depending on availability. The length of coverage varies by country from as far back as 1990 until the mid-2000s.

- OECD *Revenue Statistics:* Latin American Countries

The OECD Revenue Statistics for Latin American Countries includes data for 27 countries from 1990 to present (three of which are OECD Member Countries). The length of coverage thus provides an advantage. However, the data does not include Total Revenue, Non-Tax Revenues or Grants. Thus, it presents only total taxes and social contributions, although within these categories, there is a high level of disaggregation by relevant classifications. Given the absence of non-tax revenue, the coverage for resource revenues is somewhat less complete. Again, where available, the data is presented at the general government level.

3.2 IMF Government Finance Statistics

The IMF's Government Finance Statistics (GFS) includes data for a large number of countries, with varying length of coverage depending on the country. It further provides a high level of disaggregation (i.e., all sub-components of revenue are present) although occasionally there will be gaps. The GFS also presents data at either the general, central, or budgetary central level of government, depending on availability. There are no details of resource revenue available in the GFS.

---

[4] The 'general government' level captures all revenues reported, including from central government and others such as local, state, or social security funds. See McNabb (2017) for a comparison of how data is reported at the different levels in IMF and OECD sources.

3.3 IMF Country Reports

The IMF Country Reports display the greatest degree of heterogeneity in data reporting and thus across the criteria outlined above. The vast majority of these data come from staff reports from periodic Article IV consultations, however other IMF documents have, in the past, also provided useful sources of information (such as statistical appendices., etc.). The variability in reporting exists both across countries and within countries over time. For most countries, it is possible to use Article IV Staff reports as supplementary information to obtain a more complete overview of total government revenue since 1980—thus providing good length of coverage—but often the disaggregation of revenue data into sub-components is limited or inconsistent. The Article IV reports have, historically, been the GRD's go-to source in terms of resource revenue although as noted above the OECD now provides such information for an increasing number of countries. Often, any one country report will only provide data at either the central or general government level, depending on availability, and the reporting of revenues on social contributions is mixed.

3.4 GDP series

Data from each of the aforementioned sources is incorporated into the GRD in current local currency units (LCU). However, the main data series for the GRD presents data as a percentage of GDP. As outlined in Prichard et al. (2014), the GDP figure (denominator) typically stems from the IMF's World Economic Outlook (WEO). This data is updated quarterly and contains the most up-to-date and complete series of cross-country GDP data available. In some limited cases where the WEO does not include a figure for a certain country-year, the second preferred source is then the most recent IMF Country Report.

4   Implementation of selection criteria: overview of practical considerations

Whilst, as stated above, there is no strict ranking in terms of which source is preferred over another in the GRD, some of the source selection criteria outlined in Table 1 are usually given priority over others (although exceptions do apply). A couple of examples are illustrative:

One of the main advantages of the GRD over other cross-country sources of revenue data is the coverage of natural resource revenues. Thus, where possible, the GRD tends to include data from a source that accounted for natural resource revenue over one that did not. This is, however, conditional on some of the other criteria being fulfilled—e.g., a source that accounted for natural resource revenue but was inconsistent compared to other sources and had a short coverage in terms of years would not be included simply due to the presence of data on resource revenues.

General government data is always preferred to central government data as, by definition, it includes revenue collected by more than just the central government and thus provides a more complete picture of revenues collected in any one economy.[5] However, there are cases in which central government data may be more complete across the length and breadth criteria, and not distinctly different from general government data (e.g., in a case where a country reports the vast

---

[5] This depends on the degree of decentralisation or federalism in a given country. In India, for example, almost half of general government tax revenue is collected at the state level, thus central government data does not nearly capture the entire picture of revenues.

majority of its revenue at central and a small amount at other levels). In such cases, central government data may be preferred.

Of the criteria listed, the one which is consistently given the least weight is the availability of data on social security contributions. *Ceteris paribus*, if one source can account for social security contributions and another cannot, then the former would be preferred, however this criterion would not be preferred over any of the others listed.

A source that contains 'very long' time coverage (e.g., back to 1980) for a number of key variables—for example, total revenue and total tax, or other main aggregates—might be preferred over one with good 'breadth' of coverage for a significantly shorter time period. This is because researchers are most often concerned with studying these variables, compared to others. Whilst disaggregated data is undoubtedly useful for studying the tax mix both across countries and within countries over time, we view it as somewhat less important than the length of coverage.

5   Implementation of selection criteria: merging of sources

As a rule, the GRD typically does not merge data from central and general government level for any one country. An exception to this might be as follows:

One example might be, if some of the subcomponents of general government (e.g. income taxes) are collected entirely at the central government level. If data existed at the general government level for 1980–95 and at central government level from 1990–2020, it would be possible to verify (due to the years of overlap) which components of total revenue were collected entirely at the central government and subsequently merge the series for those components/for components where overlapping years exist and values do not differ. In most cases, the total revenue or total tax variables differ between these series if collection varies by level, in which cases we would not then not merge the entire set of revenue indicators.

This kind of merging of central and general level data would only be carried out in cases where there was (i) no better alternative, namely a consistent source of either central or general government data that spanned the entire timeframe required in the GRD, and (ii) some overlap allowing for the approach to be verified. Table 2 shows the case of the United Kingdom.

Table 2: Central and general example: OECD

| Year | General | | Central | |
	Total Tax	Income Tax	Total Tax	Income Tax
2010	514,729	144,502	388,189	144,502
2011	543,044	149,353	411,993	149,353
2012	549,944	146,474	415,694	146,474
2013	571,259	152,668	434,197	152,668
2014	597,825	156,187	456,730	156,187
2015	621,767	163,291	474,670	163,291
2016	651,251	169,582	494,543	169,582
2017	680,029	177,182	513,034	177,182
2018	705,206	184,968	530,998	184,968
2019	731,056	192,462	547,186	192,462

Note: all figures in millions of British pounds sterling.

Source: OECD (2021).

Whilst both central and general data is complete for the United Kingdom, the example serves to highlight that some sub-components of tax revenue are collected entirely by the central government—in this case, taxes on income, profits and capital gains. Thus, *were it is the case* that general government data did not provide complete coverage on income taxes, these could be merged in from central, with no loss of information.

Beyond considerations of the level of government, a more common case where sources might be merged for one country arises when two distinct sources offer different periods of coverage, or a new source becomes available that improves on the existing one across one or more of the aforementioned criteria (e.g., breadth of coverage). An example comes from Benin, where the original source in the GRD were IMF country reports, which in the 1980s and 1990s contained a fairly complete disaggregation of tax components. This was then not the case for more recent years. The IMF's GFS began reporting data for Benin in the early 2000s, where the main aggregates (e.g., total revenue) matched the series in the IMF country reports quite closely but contained better disaggregation.

In order to make use of the greater breadth of coverage offered by the GFS and merge in GFS data, two criteria would need to be fulfilled, namely:

(i)   there is some overlap between the two sources;
(ii)   where this overlap exists, we can confidently say that the sources are measuring the same concepts.

Table 3. Merging sources for Benin, central government

Benin; total government revenue, excluding grants (central) % of GDP	IMF CR 1996–2005	IMF CR 1996–2000; GFS 2001–05
1996	10.4%	10.4%
1997	10.5%	10.5%
1998	10.6%	10.6%
1999	10.4%	10.4%
2000	10.6%	10.6%
2001	10.5%	10.7%
2002	10.9%	11.2%
2003	11.1%	11.1%
2004	10.8%	11.3%
2005	11.1%	10.7%

Source: GRD (UNU-WIDER 2016); GRD (UNU-WIDER 2021).

In such a case, we can see that both sources are, broadly, measuring the same amount of total revenue, thus we can fairly confidently merge the GFS after 2001, in order to provide a better breadth of coverage.

Note: in the example above, both sources are measuring roughly the same amounts of total revenue. It is often the case that two sources do not match *exactly* but closely such that it may still be possible to merge them, provided that a period of overlap confirmed that they were following the same *trend*. Furthermore, the GRD generally prefers to include tax and revenue data from the IMF's GFS or OECD's Revenue Statistics, where available, as they are, *ceteris paribus*, relatively more complete across the breadth dimension compared to the IMF's Country Reports.

As a rule of thumb, if a source change leads to a large jump or fall in revenue figures, then the GRD would not merge two sources, as they would be measuring the underlying concepts differently.[6]

6   Exceptions

The main exception to the sources and source selection criteria outlined above applies when a national source (produced by, e.g., a statistics authority, ministry of finance, or revenue authority) produces data that fulfils three criteria, namely:

(i)   the revenue statistics are compiled according to a comparable methodology to those employed in the GRD, such as the IMF's Government Finance Statistics Manual;

(ii)   the data (e.g., revenue or tax totals) can be verified by comparing with another cross-country source;

---

[6] It may be the case that a source change happens to *coincide* with a jump or fall in total revenue; provided that a period of overlap confirmed that this was exogenous to the source change, then the GRD would change sources but include a note in order to highlight this to users.

(iii) the data contains additional information not contained in one of the usual cross-country sources, such as data on resource revenues, or subcomponents of total tax.

If these criteria are fulfilled, then the GRD could include data from such sources. For example, data for resource revenues in Egypt are currently sourced from the Egyptian Central Bank's *Monthly Statistical Bulletin.*[7] The amount of total tax in these publications is verifiable against that published in the IMF's Country Reports, but the revenues from income and profit taxes are broken down according to the source, specifically those from Suez Canal Revenues and the EGPC.[8]

One useful data source for natural resource revenues is the Extractives Industry Transparency Initiative (EITI). Annual reports from member countries classify resource tax and non-tax payments according to the IMF's GFS classification system. Thus, criteria (i) and (iii) above are fulfilled. However, the EITI reports do not contain information on the *totals* of each tax or non-tax position (including resource and non-resource revenues), meaning that criteria (ii) is not fulfilled. Hence, we cannot include the resource revenue data from EITI in the GRD, as the **non-resource revenue figures** would be computed using data from two different sources for one observation.[9]

7   Approach to data revisions

For each of the revenue sources discussed above, the GRD always incorporates the most recently produced and available data. For example, the OECD and IMF revise existing figures in the *Revenue Statistics* and *Government Finance Statistics* on an annual basis which are then incorporated during the GRD's annual update cycle. Similarly, should an IMF country report provide updated revenue data compared to a previous years' report, it would also be included in the latest GRD update. Regarding the GDP series, the periodic updates to the WEO also, crucially, reflect any rebasing that might occur by a country's statistical office. Oftentimes—especially in developing countries—when GDP is rebased, there can be significant implications for the level of a country's revenue or tax ratio. The GRD updates the underlying GDP series annually to reflect any such revisions.

References

UNU-WIDER (2016). 'Government Revenue Dataset'. Version 2016. [note: dataset no longer available]. Helsinki: UNU-WIDER.

McNabb, K. (2017). 'Toward closer cohesion of international tax statistics: The ICTD/UNU-WIDER GRD 2017'. WIDER Working Paper 184/2017. UNU-WIDER: Helsinki. https://doi.org/10.35188/UNU-WIDER/2017/410-0

OECD (2021). *Revenue Statistics*. Available at: https://stats.oecd.org (accessed 1 June 2021).

Prichard, W., A. Cobham, and A. Goodall (2014). 'The ICTD Government Revenue Dataset'. ICTD Working Paper 19. Brighton: International Centre for Tax and Development. Available at: https://www.ictd.ac/publication/the-ictd-government-revenue-dataset/ (accessed May 2021).

---

[7] https://www.cbe.org.eg/en/EconomicResearch/Publications/Pages/MonthlyStatisticaclBulletin.aspx

[8] Egyptian General Petroleum Corporation

[9] That said, we do include estimates of resource revenue from EITI in the notes of the GRD, for reference purposes.

UNU-WIDER (2021). 'Government Revenue Dataset'. Version 2021. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/GRD-2021