# Do Performance Measures of Donors' Aid Allocation Underperform?

*Paul Clist*
*School of International Development, University of East Anglia, Norwich, UK*

## 1. INTRODUCTION

ON 13 August 2006, the Center for Global Development released that year's commitment to development index (CDI), an attempt to rank donors by their overall contribution to development, placing Japan last in 21st place. Within a month, the Japanese Ministry of Foreign Affairs released a statement arguing that 'this Index neglects the strengths of Japanese aid, drawing on a certain value judgement' (MOFA, 2006). This incident could be interpreted as highlighting CDIs ineffectiveness in the task of influencing policy, given that their judgement was so summarily dismissed. Alternatively, the fact that Japan engaged with the measure so quickly could constitute evidence of its success, showing the power of an index to stimulate debate. In the intervening years, the number and profile of rankings have grown. This increase has taken place in the context of a desire to increase the effectiveness of foreign aid, given the lack of robust evidence of any growth-promoting aid effect (Roodman, 2007). Clist (2011) has shown that bilateral donors are heterogeneous, and Kilby and Dreher (2010) argue that this causes a differential aid effect. Collier and Dollar (2002) even claim that aid-induced poverty reduction could double as a result of allocating aid differently. If gains of this magnitude are possible, it is clear that even partially successful political pressure on donors could reap large dividends. This makes change at the donor level very promising, and rankings are often designed for this task. Thus, to their supporters, rankings offer not only a useful judgement on current practice, but more importantly represent a mechanism for improving aid effectiveness. To their detractors, rankings simply draw on 'a certain value judgement', and so are rightly dismissed.

The role and validity of donor performance measures are examined in light of their stated aims and growing popularity. In a thoughtful essay, Roodman (2011b, p. 483) admits that composite indices 'are inherently crass' but argues that this is justifiable 'as long as the aggregation methods and implied valuations are clear, and there is honesty about the goal of public communication'. Høyland et al. (2012) argue that one way to improve indices (they focus on data from Doing Business, the human development index and Freedom House) is to incorporate a measure of uncertainty into the index, to give a better understanding of the level of uncertainty and measurement error. A competing viewpoint is that indices should not be improved but abandoned, as they are not just crass but unhelpful. This viewpoint has been expressed in the context of poverty measures, where these issues have been debated more thoroughly. Ravallion (2011) questions the very need for aggregation arguing that a range of

disaggregated measures is a preferred alternative. This scepticism has a long tradition; McGillivray (1991) asked if the human development index was 'yet another redundant composite development indicator'. It is interesting to note that this author was concurrently proposing an index of donor performance. This highlights the extent to which the literature on donor performance measures has run in parallel to the more general critique of composite indicators. This paper is an attempt to remedy this by bringing relevant insights to bear on influential rankings and indices.

I focus on measures of how desirably a given amount of aid is allocated among potential recipients (donor allocative performance). This is not to say other aspects of aid policy are unimportant, rather this focus is chosen as one of the few aspects that a number of indices deal with, allowing for a comparison of alternative approaches. The plurality of measures is easily explained by the relative importance and ease of measurement of allocative performance. In Section 2, I seek to make clear the number of decisions that are taken in designing an index of donor performance, focusing on five measures. The section discusses three simple choices: what to measure, how to measure and how to aggregate. Section 3 presents some simple sensitivity tests, showing how measures change when various methodological choices are varied. Only minimal changes are made, and yet each change leads to different rankings. Section 4 presents results from these five measures, and shows that their judgements overlap very little. I argue that this threatens to undermine the aim of descriptive indices, as political pressure is difficult to sustain in the face of contradictory and competing judgements that are not robust to minimal methodological changes. Section 5 discusses solutions to this problem, presenting a graphical alternative which avoids the major problems of the descriptive approach. Section 6 concludes.

## 2. THE DESCRIPTIVE APPROACH

I focus on five performance measures of donor allocative performance, chosen as representative of the most influential and/or recent approaches. Some of these are a more general measure of which I discuss only the subcomponent that deals with selectivity. First, the MPI (McGillivray performance index; McGillivray, 1989) is an index that starts by attaching a weight to each potential recipient according to its desirability. As the weights fit within the range 0 to 1, they can be thought of as discounts applied to aid given to less desirable recipients. Aid to the richest country is completely disregarded (i.e. it is given a weight of 0), and aid to the poorest country is not discounted (i.e. a weight of 1). The final index is then the sum of weighted aid divided by all aid, scaled to fit between 0 and 100. If all aid were given to the poorest recipient, then the donor would receive 100: the best possible score. Second, the API (adjusted performance index) is McGillivray's (1992) attempt to respond to criticisms of the MPI regarding how it includes population. Specifically, as aid *per capita* is used in the MPI, the perverse situation can arrive where a reallocation from a large poor country to a small rich country can improve performance (White, 1992). This is because a set amount of aid is larger in *per capita* terms when given to less populous countries. To combat this, the API uses population divided by income as the weight, with the maximum score now obtained by giving to the recipient with the largest population/income score. The third, RPI (Roodman performance index) is the selectivity weight used in the wider CDI discussed in the introduction. It is perhaps the most high-profile index discussed here, as it is used explicitly by the Dutch and Finnish governments, has influenced Australian, Canadian and Norwegian policy, angered Japanese officials and is supported financially by ten bilateral donors (Roodman,

2006). The RPI resembles both the MPI and API, but policy[1] replaces population as a factor of interest. Population does influence another part of the broader index in determining proliferation, but is not incorporated into the measure of selectivity at all. The policy and poverty weights are found separately in a similar fashion to the MPI and API, and then multiplied together to obtain the weight used in the index. The fourth, EW (Easterly and Williamson, 2011) does not sit within the traditional index method, but rather uses the headcount measure. This means that it calculates the percentage of a donor's aid that meets a given criterion. Three variables are chosen and given additive weights: low-income countries (50 per cent), politically free countries (25 per cent, based on Polity VI data) and less corrupt countries (25 per cent, based on ICRG data). Thus any aid to a country that meets all three criteria receives a weight of 1, with weights of 0.75, 0.5, 0.25 and 0 for other countries, depending on which criteria they meet. The fifth measure, KRE (Knack et al., 2011) uses the average of three coefficient estimates from a sparsely specified regression: aid is regressed on income, policy[2] and population (where all variables are logged). The coefficient estimates are then combined to give a final score.

### a. The Aim and Approach of Rankings

Measures of donor allocative behaviour typically aim to change the behaviour they measure (Easterly and Williamson, 2011; Knack et al., 2011). Birdsall (2011) states three ways in which this might happen: by changing the focus of a debate, by highlighting certain technical issues and by encouraging advocacy through public communication. These three ways clearly overlap, and can be subsumed under the general heading of policy influence. Roodman (2011b) states that their main use is in educating the public, which is the first step in creating political pressure for a change in behaviour. For this purpose, rankings offer clear advantages: they condense a large amount of information into an easily understood format. Furthermore, they are media-friendly sources of national pride, shame and controversy. This feeds into the aim of changing donor behaviour through influencing public opinion.

The aim of public education explains not only the relatively simple methods of presentation, but also the preference for simple methods of measurement. The selection of simple, easy-to-understand methods coincides with a growing distrust of regressions (Roodman, 2007). This distinguishes the descriptive research from explanatory research (e.g. Alesina and Dollar, 2000; Berthélemy and Tichit, 2004; Clist, 2011), which seeks to explain allocations by fully specifying a regression. By contrast, the descriptive approach merely describes a certain aspect of an allocation, meaning they are more *ad hoc* and flexible in their approach. In avoiding the need to fully specify regressions, descriptive measures often claim to be methodologically simpler and by implication more trustworthy: this claim can be found in Roodman (2004; p. 18) a descriptive approach 'minimizes questions about proper modeling specification', Nunnenkamp and Thiele (2006; p. 1179) 'we follow Roodman . . . who stresses the risk that cross-country regression models are misspecified and, thus, favours a simpler approach', and Easterly and Williamson (2011, p. 2)'Once the Pandora's box of conditioning factors is opened, it is very hard to decide where to begin or where to stop'. However, this claim is not quite as well established as it is widely repeated. While it is clear that these approaches do

---

[1] Specifically, this is an average of the six measures of the Worldwide Governance Indicators.
[2] The variable used is the country policy and institutional assessment (CPIA) score.

not claim to know the data generating process, this does not automatically equate with minimising questions of proper specification. It merely means choosing a different set of questions: it is with those questions that this section deals.

### b. What to Measure

The variables chosen in any measure seek to reflect a factor that is thought to constitute part of the latent variable under examination. In many cases, the latent variable is rather grand and difficult to define, but here it is donor allocative performance: the desirability of a distribution of a given amount of aid among a given set of potential recipients. Some of the indices I discuss have a broader focus, of which I examine merely the selectivity subcomponent. Having decided the latent variable, there are two questions related to what to measure: the factors and variables. Which factors make up allocative performance is surprisingly contentious. Clist (2011) introduced the 4P framework (poverty, population, policy and proximity) to explain aid allocation, the first three of which are the most commonly mentioned factors in normative accounts (Llavador and Roemer, 2001; Collier and Dollar, 2002). Poverty has the broadest appeal, and is typically measured using income *per capita* given incomplete poverty data. This proxies for both the amount of poverty in a country and the resources for dealing with this poverty. All five measures discussed here include it. Population is a measure both of the amount of poverty (in combination with a measure of income *per capita*) and potentially the marginal effectiveness of aid.[3] Most normative accounts discuss population and most measures include it; however, the RPI and EW are completely insensitive to population in the selectivity part. This leads to some surprising implications; insensitivity to population means that all donors could improve their RPI selectivity score by reallocating all of their aid to the Pacific island of Kiribati (with a population of around 100,000).

The idea that the policy environment of a country influences the marginal effectiveness of aid is widely recognised due to Collier and Dollar (2002), but also hotly contested. An alternative rationale for allocating aid in response to policy is to incentivise certain policy environments. While the RPI, EW and KRE include policy, the API and MPI do not. In both cases the exclusion of policy is explained as the direct result of a normative belief that aid should be allocated on income (and population) grounds alone (White and McGillivray, 1995). However, the selection of what to measure is influenced by a normative vision in every case, be it explicit or implicit. All measures of donor allocative performance will necessarily contain a judgement as to the correct factors to include, which will reflect certain ideals and beliefs. While these three factors (Poverty, Population and Policy) are the most commonly discussed, none of the five measures are always sensitive to all of them (Anderson and Clist, 2011). This is due to both deliberate decisions resulting from normative differences and technical features of the measures.

It is worth noting that measures of policy or governance differ from measures of income and population in that they are necessarily more subjective: there is disagreement over the underlying factor not just the specific variable. While there may be technical disagreements regarding the correct measure of income or population, there is greater room for philosophical and political disagreements regarding what constitutes a good policy or institutional environ-

---

[3] If there are fixed costs or economies of scale in providing aid to a given country, then population would be one factor that influences the marginal effectiveness of aid.

ment. Disagreement on the donor side is obvious, for example, there are clear differences of opinion between the United States and Nordic donors over the correct type of regulation in an industry. Between researchers creating indices, this disagreement can be more subtle. For example, Easterly and Williamson (2011) and Knack et al. (2011) both purport to measure policy selectivity. To represent this, the latter use the World Bank's in-house measure of policy (CPIA) that emphasises public sector and economic management, whereas the former are more focused on political freedom and corruption. The difference in variable choice reflects a more fundamental disagreement. Both sets of authors argue that they are measuring selectivity, despite very different theoretical conceptualisations of what constitutes policy selectivity.

### c. How to Measure

One dividing line between explanatory and descriptive research is that descriptive research typically measures absolute characteristics, whereas explanatory research measures conditional characteristics. The desirability of controlling for other factors depends upon the goal of the research. Knack et al. (2011) argue, from within the descriptive tradition, that controlling for other factors is useful if the judgements are to be used to influence policy, as it controls for the limitations on donor actions. However, the approach of KRE appears flawed as they control for only some aspects of allocation but leave out important confounding factors such as historical, linguistic and commercial links between recipient and donor. As such, it cannot be thought of as a fully specified regression, but neither does it constitute a test that is always sensitive to the three factors it measures (Anderson and Clist, 2011). In this way, the approach of KRE appears to miss out on the absolute advantages of both the descriptive and explanatory approaches. Easterly and Williamson (2011) argue that examining absolute sensitivity provides an informative alternative, complementing explanatory work. This seems a sensible role for explanatory research, as while the positive literature is more econometrically advanced, it might be that controlling for other factors masks population, policy and income sensitivity in absolute terms. Imagine that the positive literature finds that a given donor has low income-sensitivity but high concern for former colonies: clearly a donor's willingness to support its former colonies may be the best explanation of its aid allocation. However, if these colonies are also poor, the positive coefficient on colonies will dampen the coefficient on income – which may lead the reader to underestimate the absolute poverty-selectivity of the donor. If donors are to decrease fragmentation they should not be punished for allowing colonial and linguistic links to influence allocation patterns, as this promises to reduce proliferation and therefore transaction costs. This makes absolute judgements of poverty sensitivity a useful tool in judging donors, as conditional judgements of poverty-sensitivity may mislead. Any conditional effect is clearly best estimated within the explanatory approach, and the descriptive approach should then concentrate on its advantage of absolute measures.

Klugman et al. (2011) discuss the importance of scale invariance within the context of the human development index (HDI). If a measure is not scale invariant, the way it is measured (e.g. using a five point or 10-point policy scale) will affect the rankings. The old HDI formula, like the MPI, API and RPI, tries to avoid this problem by rescaling the index with reference to minima and maxima. However, this introduces the problem of sensitivity to those bounds. This is then another source of 'methodological uncertainty' that some purport to minimise by using the descriptive approach. The most common index approach to scaling data is to use the form $x = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}}$: this is found in the RPI and MPI. This attaches a weight ($x$) to

the recipient based on its desirability as an aid recipient. To understand the effect of this, consider the weight for the three common factors (population, policy and income) for the median potential recipient in a hypothetical index. For population, using 2008 data, $N_{max}$ is 1.3 billion (China), $N_{min}$, is 31,000 (Palau), and so the median recipient (Laos, 6.3 million) has a section weight of 0.005 (calculated using logged population, where China would have a weight of 1 and Palau 0). For income, Guatemala is the median recipient with an income of US$ 4,285 (*per capita*): the section weight is 0.14 (using logged income data). For policy (WGI), the median recipient is Malawi with −0.33: the section weight is 0.45. We can think of these as discounts of 99.5, 86 and 55 per cent, respectively. Therefore, because the typical index scaling method relies on minima and maxima, and population is heavily positively skewed, any aid to the median recipient in terms of population is essentially dismissed.[4]

### d. How to Aggregate

Once the different aspects of allocative performance have been decided upon and measured, descriptive approaches commonly aggregate them to create a ranking based on a single number. While in theory the different factors can be weighted, an equal weighting is often used. This is chosen more as a default rather than for any inherent attractiveness or theoretical justification. Ravallion (2012, p. 15) argues that the 'degree of robustness to weights depends on the intercorrelations among the components. If these are perfectly correlated then … the result is entirely robust to the choice of weights'. Figure 1 is a scatter graph of poverty and policy for potential recipients of aid. It shows that the two variables are very highly correlated, but this correlation is negative (the correlation coefficient between income and policy is 0.72). Thus the relative implicit weight given to the two factors will greatly affect any final rankings. The weight is not just the statistical parameter that is called weight, but rather the sensitivity of rankings to changes in that part of the data. From the example above, it is clear that the index method of rescaling the data gives greater weight to population even if this is aggregated using an 'equal' weighting system.

Two approaches to combining the various factors are common: geometric and arithmetic averages. For example, the geometric average is used by RPI, and an arithmetic average is used by EW. The simple difference between the two approaches is that under the geometric approach, donors are punished more for simultaneously performing badly in multiple areas. In contrast, under the arithmetic average, a given portion of the overall weight is decided by a specific factor: for example, EW give 50 per cent of the weight to income and 50 per cent to policy. To illustrate the difference between the two approaches, imagine a recipient exhibited the median characteristics described above. Using a geometric average, any aid would receive a weight of 0.000315, whereas under an arithmetic average it would receive 0.199. Thus, aid given to this imaginary median recipient would either be discounted by 99.99 per cent or 80.1 per cent. The heavily skewed nature of population determines a third of the arithmetic weight, but completely dominates the geometric average. A further problem with the geometric average arises when the individual components are allowed to be negative. This could potentially happen with the RPI, as the minima and maxima are set in relation to data from the first year of its calculation. Thus, a recipient with either a lower governance score than Afghanistan (in

---

[4] Regression coefficients are not directly comparable, but the strong effect of outliers is widely recognised. To summarise briefly the difference, they do not use maxima and minima as their points of reference but deviations from mean values.

FIGURE 1
Scatter of GDP *per capita* and the WGI, 2009
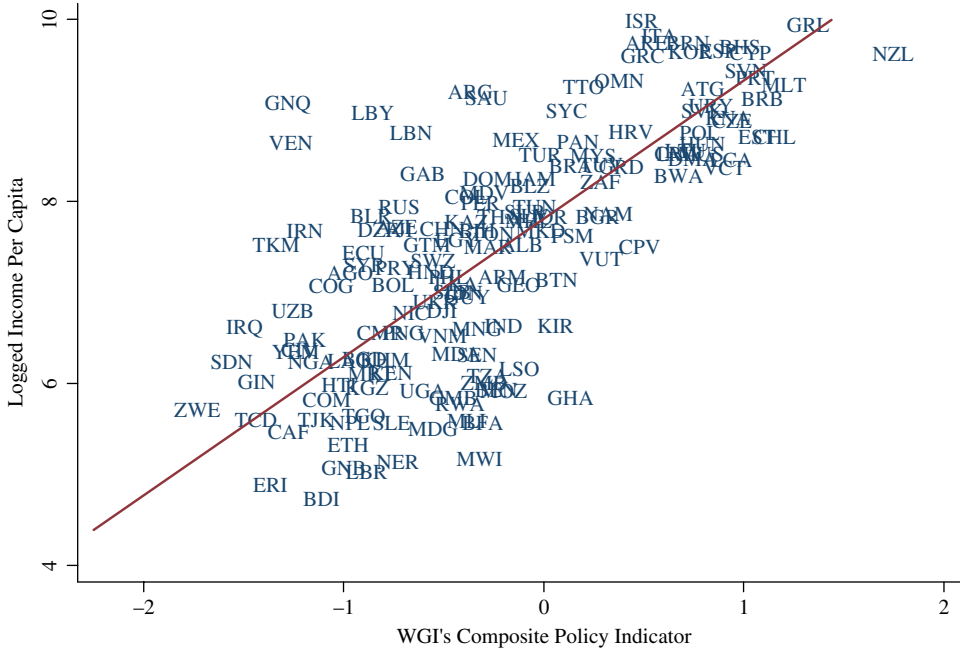


Note:
ISO names used as markers, the line denotes the maxima and minima from the RPI.

2000) or higher *per capita* income than Singapore (in 2001) will receive a negative score. This is troubling as a slight change in one variable could have a large effect on the index. While unlikely in the case of the RPI, this approach can give perverse results if two components are negative. In this case, a recipient would receive a high weight because of its exceptionally poor characteristics.

## 3. SOME SIMPLE TESTS OF ROBUSTNESS

The preceding section separates out the decisions of what to measure, how to measure and how to aggregate. However, the effect of each decision needs to be understood in conjunction with the others'. The relative importance of one factor is not determined solely by how factors are aggregated, but is also influenced by decisions over variables and measurement. This section examines what effect these decisions have on final rankings, by varying some of these decisions. It is not a full sensitivity analysis, as the number of possible descriptive measures makes this impractical (a case of combinatorial explosion). Instead, I show the effect of simple, minimal diversions from a measure's implementation, beginning with the EW and KRE measures. EW uses a headcount method, and so trade-offs between factors are quite easy to understand. EW calculate the amount of aid going to low income (using a World Bank classification), free (indicated by a Polity IV democracy score of 8 to 10) and less corrupt countries

(from the ICRG dataset). This could be viewed as an index method, where potential recipients receive a weight of 0, 0.25, 0.5, 0.75 or 1. There is no change to the weight of a recipient unless it crosses a boundary: two recipients that have polity scores of 1 and 7 are equally dismissed, and recipients with polity scores of 8 and 10 are equally valued. In the same way, all middle-income countries are discounted, and all low-income countries valued. Unfortunately, given the proprietary nature of the ICRG data used, I am not able to calculate the effect of small changes in methodology. In lieu of this, I discuss the specific problems and anomalies of the headcount approach in this setting.

Table 1 lists the relevant characteristics of six countries, along with the implied weight this recipient would receive (this abstracts from the ICRG corruption data, and so 0.25 is left off all recipients). This illustrates the problem of thresholds in the headcount approach, where large differences in characteristics can mean no difference in the weight, and yet conversely, small differences that cross a threshold can mean very different weights. So, Senegal and Argentina are equally dismissed as non-low-income countries whereas Guinea-Bissau and Kenya are equally valued as low-income countries. Only Argentina and Ghana are valued for their Polity IV score, with all others receiving the same discount for their low scores. The table also illustrates a specific problem with the headcount measure in this instance. As shown in Figure 1, income and policy data is typically highly positively correlated. Because of this, a headcount measure is more likely to value recipients that just meet a given criterion. In other words, of all recipients that meet the low income criterion, the relatively richer are the most likely to meet the policy criteria.

What effect does varying a threshold have? I investigate the effect of changes in the threshold that determines a country's classification as 'free', as a test of sensitivity of EW rankings to a minimal methodological change. Easterly and Williamson (2011) use Polity VI data, where a score of 8 or above is classified as free. I calculate the effect of varying this threshold to be 7 or above, with original and alternative rankings shown in columns 1 and 2 of Table 2. These rankings only refer to the freedom part of their selectivity measure. There are substantial rises in the rankings for Austria, Japan and the UK, and large falls for Luxembourg, Sweden and the USA. Only four donors do not move, and the average change in ranking is 2.8 places; the judgement of donor sensitivity to democratic principles is sensitive to threshold choice. This choice is arbitrary, with little justification or guidance for choosing one number over another. This sensitivity is perhaps to be expected with a headcount approach. However, because of the negative correlation between income and governance data (Easterly and Williamson, 2011) the threshold choices appear crucial.

The KRE differs from the other measures as it uses a regression to determine the scores of donors. The final score is an arithmetic average of the three coefficient estimates, and as the

TABLE 1
EW Weights for Selected Countries

| Recipient | GDP per capita | Low Income | Polity IV | Implied EW Weight |
|---|---|---|---|---|
| Senegal | 1,656 | No | 7 | 0 |
| Argentina | 13,220 | No | 8 | 0.25 |
| Guinea-Bissau | 966 | Yes | 6 | 0.5 |
| Kenya | 1,429 | Yes | 7 | 0.5 |
| Uganda | 1,067 | Yes | 1 | 0.5 |
| Ghana | 1,375 | Yes | 8 | 0.75 |

TABLE 2
EW and KRE Sensitivity of Bilateral Donor Rankings for 2009

| Donor | EW Orig. Free if ≥8 | EW Alt. Free if ≥7 | KRE Orig. CPIA | KRE Alt. WGI |
|---|---|---|---|---|
| Australia | 1 | 1 | 22 | 20 |
| Austria | 15 | 7 | 15 | 11 |
| Belgium | 9 | 11 | 7 | 7 |
| Canada | 19 | 23 | 9 | 6 |
| Denmark | 21 | 22 | 12 | 8 |
| Finland | 13 | 17 | 16 | 10 |
| France | 11 | 9 | 18 | 19 |
| Germany | 14 | 10 | 2 | 3 |
| Greece | 22 | 20 | 20 | 23 |
| Ireland | 4 | 5 | 4 | 5 |
| Italy | 10 | 8 | 19 | 18 |
| Japan | 23 | 18 | 5 | 2 |
| Korea | 5 | 4 | 11 | 17 |
| Luxembourg | 12 | 19 | 14 | 14 |
| Netherlands | 18 | 21 | 13 | 15 |
| New Zealand | 2 | 2 | 21 | 21 |
| Norway | 6 | 6 | 6 | 12 |
| Portugal | 3 | 3 | 23 | 22 |
| Spain | 17 | 15 | 8 | 9 |
| Sweden | 8 | 12 | 10 | 16 |
| Switzerland | 16 | 14 | 17 | 13 |
| UK | 20 | 16 | 3 | 1 |
| USA | 7 | 13 | 1 | 4 |
| Average change | | 2.78 | | 2.61 |

Note:
(i) Orig. and Alt. denotes rankings by the original and alternative methodologies. EW rankings are merely on the aid to 'Free' countries section.
(ii) Average change refers to the average number of places difference between the original and different methodologies.

three coefficients are determined simultaneously, the relative weight given to each factor varies by regression and donor. To test the sensitivity of KRE to methodological differences, I recalculated it using the WGI as the policy indicator, as opposed to the CPIA.[5] No other methodological points were changed, and the subsequent rankings of the donors can be seen in column 4 of Table 2 alongside the original ranking in column 3. Only three donors have the same ranking (Belgium, Luxembourg and New Zealand), and the average change is 2.6 places. Korea and Norway appear to be much worse donors when using the WGI data rather than the CPIA, dropping from 11th to 17th and from 6th to 12th, respectively. Conversely, Finland is viewed more favourably, climbing from 16th to 10th. These differences could be the result of differences in samples, as the coverage of the two variables differs. The WGI and CPIA share very similar aims, the former even incorporating the latter in its assessment. For the donors that look worse because the CPIA is used, the WGI is clearly preferable. For others however, it is difficult to see how an objective decision could be made, and how one should choose between the two sets of results is far from clear.

[5] As the WGI variable ranges from −2.5 to 2.5, its log was calculated after adding 3.5.

Table 3 deals with alterations to the RPI. I take the RPI as representative of indices, as the API and MPI do not have a specific and detailed canonical method of implementation (e.g. there is no guidance on selecting the sample). While specific details may differ,[6] these arguments apply to the API and MPI in general terms. The marginal rate of substitution (MRS) is a useful concept in understanding the three indices. It describes the amount of one variable that must be increased to compensate for a shortfall in another. The index method of rescaling the data means that the maxima and minima in each variable are crucial as they determine the MRS. The line in Figure 1 represents the MRS used in the RPI, running from the maximum to the minimum. It can be thought of as an isoquant, where any recipients lying on the line are equally valued. The best recipient would reside in the south east corner of the graph, and be both poor and well governed. For GDP *per capita,* the bounds are $21,869 and $81 given by the 2001 scores of Singapore and the Democratic Republic of Congo, respectively. For the governance score, the range is −2.25 to 1.44, given by the scores of Afghanistan and Singapore (2000 data was used as 2001 data do not exist). The RPI equates the two scales: thus $21,788 *per capita* dollars (logged) are equated to 3.69 on the WGI policy score. Using a different sample, variable or base year would change the minima and or maxima, which would in turn change the MRS and ultimately the relative performance of donors. I examine each of these methodological choices in turn.

The choice of sample may affect the MRS through deciding the maxima and minima of the measure. In the RPI, a number of countries are excluded as being rich[7] in a decision made by the researchers in the first year of the CDIs operation, on the basis of which countries were plausible aid recipients. However, Israel is excluded despite being a very large aid recipient. Neither is it a strict implementation of an income threshold in 2002, as countries of similar incomes to those excluded are left in. For example, Portugal (21,372) is excluded as rich whereas Singapore[8] (36,076) and Cyprus (23,590) are included. Singapore, a high income country, determines one end of both policy and income scales, and the inclusion of these high income countries leads to some surprising implications. For example, Poland received a selectivity weight of 0.47 as a donor in 2009 (Roodman, 2011a), but in that same year it received a score of 0.59 as a potential recipient (Roodman, 2011b). Thus, by ceasing to allocate aid to other countries and instead giving all its aid to itself, Poland would improve its selectivity score by around 25 per cent. This is not an isolated case as several of the donors that are included in the sample would receive a higher selectivity score if they reallocated all their aid to themselves: the Czech Republic would increase from 0.58 to 0.59, Turkey from 0.40 to 0.47 and Hungary from 0.57 to 0.61.

While the RPI methodology does not change and the base year is fixed, it updates every year such that new data in 2001 is allowed (if applicable) to update the reference points of maxima and minima. The sample of excluded countries does not include Bermuda, which is included as a potential recipient for 2009 data, despite having a GDP *per capita* of 66,268

---

[6] Most notably, the RPI includes policy and not population as a factor, with the associated problem of greater subjectivity.
[7] Specifically, they are Austria, Belgium, Denmark, France, Germany, Italy, Netherlands, Norwich, Portugal, Sweden, Switzerland, UK, Finland, Iceland, Ireland, Luxembourg, Greece, Spain, Canada, USA, Israel, UAE, Japan, Taiwan and Australia. Thanks are due to David Roodman for clarifying this, and several other points (personal correspondence).
[8] In documentation of the RPI, Singapore's 2002 GDP *per capita* is stated to have been 21,869. However, current data show it as 36,076. It is not clear whether Singapore is actually included in the base year, where this figure comes from, nor what effect it has had, if any.

TABLE 3
RPI Sensitivity of Bilateral Donor Rankings for 2009

| Donor | RPI Original | Different Sample Including Bermuda | Different Variable | Different Year | |
|---|---|---|---|---|---|
| | | | | Highest | Lowest |
| Australia | 14 | 17 | 16 | 16 | 17 |
| Austria | 15 | 13 | 14 | 12 | 14 |
| Belgium | 16 | 18 | 13 | 18 | 18 |
| Canada | 9 | 10 | 7 | 8 | 10 |
| Denmark | 1 | 1 | 2 | 1 | 1 |
| Finland | 7 | 7 | 15 | 7 | 8 |
| France | 22 | 19 | 19 | 19 | 21 |
| Germany | 18 | 14 | 5 | 13 | 15 |
| Greece | 23 | 22 | 4 | 19 | 23 |
| Ireland | 3 | 3 | 20 | 3 | 4 |
| Italy | 20 | 23 | 6 | 20 | 23 |
| Japan | 11 | 9 | 10 | 8 | 10 |
| Korea | 6 | 5 | 21 | 4 | 5 |
| Luxembourg | 2 | 2 | 1 | 2 | 2 |
| Netherlands | 13 | 16 | 11 | 13 | 17 |
| New Zealand | 21 | 20 | 22 | 21 | 23 |
| Norway | 12 | 12 | 17 | 11 | 12 |
| Portugal | 4 | 4 | 23 | 3 | 5 |
| Spain | 17 | 15 | 3 | 15 | 17 |
| Sweden | 10 | 11 | 18 | 11 | 13 |
| Switzerland | 5 | 6 | 9 | 6 | 6 |
| UK | 8 | 8 | 8 | 7 | 9 |
| USA | 19 | 21 | 12 | 19 | 22 |
| Average change | | 1.39 | 6.96 | 1.39 | 1.43 |

Note:
(i) The RPI original column was calculated using the maxima and minima reported in Roodman (2011b).
(ii) In columns 4 and 5, as well as Figure 3, I use the actual data and so rankings differ (see footnote 5 for one such difference).
(iii) Average change refers to the average number of places difference between the original and different methodologies.

(Roodman, 2011a).[9] Currently missing data for other variables in 2002 mean that Bermuda does not set the income maxima in the base year (with 59,699). If the data were provided, it would have a large effect on the MRS, with 3.32 on the governance scale equating 59,618 on the income scale. Table 3 shows the rankings using original methodology in column 1, and those allowing Bermuda to set the maximum income in column 2. Of the 23 donors, seven are unchanged, with an average change in ranking of 1.4 places. The seemingly small decision to include Bermuda means Italy and Australia fall by three places and France and Germany to rise by 3 and 4 places, respectively. These changes are not the result of a large change in sample, and Bermuda's exclusion appears to be the result of data coverage rather than a deliberate decision. Thus, even this small change of sample affects the rank of the majority of bilateral donors. The decision of which recipients to include is a small detail, was

---

[9] This does not receive a negative selectivity weight because the log of income is used.

taken without any theoretical guidance and was not the strict application of a threshold. I do not wish to argue that the RPI made the wrong choice in selecting the sample, but rather a more pessimistic point: there is no obviously superior way of deciding which countries in the world are potential recipients, and yet this decision has real effects in determining the MRS and the subsequent rankings of donors.

Column 3 of Table 3 reports the alternate rankings found if the Polity VI variable used by EW is used instead of the WGI. As aforementioned, changing the variable changes the sample due to differences in data coverage. Thus, it is to be expected that the measure is more sensitive to the change in variable to allowing one extra data point (the Bahamas in 2002). Even taking this into account, the differences are striking: an average change of seven places.[10] Greece and Portugal switch places due to the change in variable: between 4th and 23rd. Large changes can also be seen in the rankings of Ireland, Korea, Italy and Spain. If the Polity VI data were used instead of the WGI, Greece, Italy and Spain would see criticism replaced with praise, with the opposite effect for Ireland, Portugal and Korea.

Turning to the base year, if 2002 were chosen instead of 2000, the bottom of the policy range would be Iraq ($-1.88$), and the top of the range would be unchanged (given by Singapore with 1.44). Assuming the income range was stable, this would equate $21,788 *per capita* dollars (logged) with 3.32 on the WGI policy scale. To investigate the effect of such changes, I recalculated the scores of each donor using base years between 1996 and 2008: columns 4 and 5 of Table 3 report the highest and lowest rank of the bilateral donors. The RPI uses 2001 as the base year as it was the most recent data available in the year the index started, meaning this exercise can be thought of as a test of how robust the rankings are to the year the commitment to development index started in. Four donor rankings are unaffected by such a change: Belgium, Denmark, Luxembourg and Switzerland. For the remaining 19, their performance relative to other donors' changes depends on which of the 13 years is used as the reference point. The changes may not seem large: the largest difference is four places, and the average change is 1.4 places. However, remember that nothing has changed apart from the base year. The choice of base year determines which of New Zealand, Italy and Greece was named the worst donor in terms of selectivity, keeping all other methodological aspects constant. These changes come from changes in the MRS that result from different minima and maxima being used. The sensitivity of rankings to the whimsical choice of the base year is worrying as even the seemingly innocuous decision of base year alters rankings. The cumulative effect of multiple small decisions can only be larger, which questions the robustness of the five approaches to even minimal changes in methodology.

## 4. DISPARITY OF RANKINGS

Despite the pessimism of Section 3, it is possible that the sensitivity of donor allocative performance measures are academic. If the popular measures concur, then any methodological difference is perhaps not overly troublesome. For this reason, I recalculate the five measures and display the results graphically, following Høyland et al. (2012). However, while they display results of sensitivity tests, Figure 2 shows the (rescaled) raw scores and rankings using the canonical methodology of the five measures. They were calculated using recent data (2010 for aid, 2009 for all independent variables) for the 23 OECD/DAC bilateral donors.

---

[10] Note that a completely reversed ranking of 23 donors would have an average change of 11.5 places.

FIGURE 2
The Rescaled Raw Scores and Rankings of Bilateral Donors, 2009
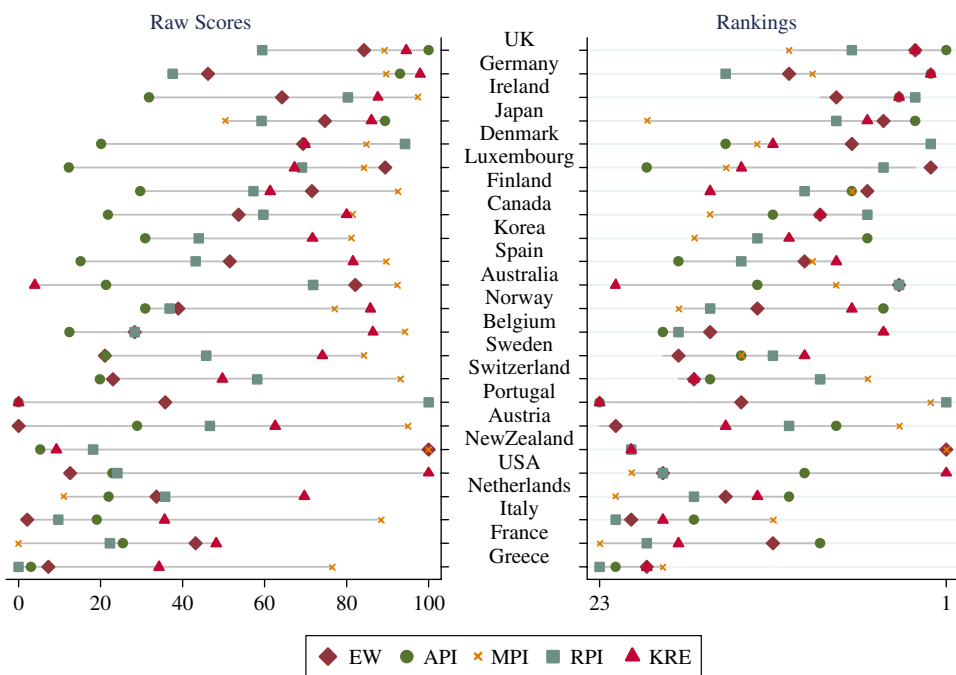


The exception is Easterly and Williamson (2011) who use a proprietary dataset for governance, so their original raw scores and rankings are used for the 22 donors that they include (they exclude Korea) using 2008 data. The left of the Figure shows the raw scores, which are rescaled to fit between 0 (the worst donor) and 100 (the best donor). The right of the figure shows the actual rankings from 23 to 1 (the actual scores are shown in Table A1 in the Appendix). This is not a full sensitivity analysis, but rather a subset of possible judgements. If more indices were included or the full gamut of possible rankings were explored, we would expect the range to increase dramatically. It is not a sensitivity test and is more likely to produce false positives than false negatives, so a failure to show an agreement on this simple test would be particularly worrying.

In Figure 2, the donors are ranked in descending order of average (rescaled) donor performance: the UK is considered to be the best donor, and the worst is Greece. If the measures tended to concur, there would be a clear diagonal line running from top right to bottom left in both raw scores and rankings. Instead, we see a great disparity of rankings. The two donors with the smallest variability are Greece and Ireland, for all others there appears to be real disagreement about how good a donor's allocation is. The best single donor according to the RPI, MPI, and KRE and EW measures (Portugal, New Zealand and the USA, respectively) are ranked, on average, among the worst. For the majority of donors, it is not clear which half of the distribution they should reside in, and there are few consistent orderings of one donor over another. It is worth stressing that this is not a test of the effect of the full range of methodological uncertainty: just the results of five popular/recent measures. The disparity of

descriptive results is not a recent trend – White and McGillivray (1995) proposed two measures that met their criteria which gave results that had a rank correlation of zero.

### a. Undermined by Their Own Flexibility

The aim of the various indices is to change the behaviour they measure. It is often envisaged that this acts through education of the public, which in turn leads to political pressure. Alternatively, some see indices highlighting certain technical matters to the governments themselves. Regardless of the mechanism, the aim is undermined by the disparity of rankings: a mass of contradictory judgements offers neither clear technical advice[11] nor consistent political pressure. The preference for methodological simplicity, in the hope that this minimises methodological uncertainty, seems somewhat misguided. The sensitivity of measures and the disparity of judgements are the natural consequence of methodological uncertainty within the descriptive tradition. The choices of what to include, how to measure these factors and aggregate these findings are important decisions which greatly affect the final rankings, and must be taken with little theoretical direction. It is hard to believe that the goal of advocacy through increased public awareness would survive a debate on methodological differences of indices, with popular opinion settling upon a preferred index. The result is a multitude of fragile and contradictory judgements which are all potentially valid, their differences stemming from justifiable choices of what to include, how to measure those variables and the aggregation methodology. Today, a donor that is ranked poorly by one measure need not state it is based on value judgements; it could merely point to its high scores from a different measure.

## 5. FUTURE DIRECTIONS

The evidence presented here is not positive regarding indices – the rankings they produce are neither robust nor in agreement. This section discusses possible future directions in the light of this evidence. First, I discuss whether more aggregation might be a viable way forward. Second, I discuss the opposite direction – whether disaggregation might sidestep the most intractable problems of the descriptive approach. Third, I question the value of the approach itself, asking whether the approach has anything to offer despite its limitations.

### a. More Aggregation

Høyland et al. (2012) follow the Worldwide Governance Indicators and seek to resolve the problem of aggregation by resorting to more aggregation, in quality and quantity. Thus, rather than aggregating merely the point estimates of different measures, they seek to aggregate the range of estimates. This allows a statement regarding the sensitivity to certain assumptions. I do not choose this route. Section 3 shows that indices are not robust to even small changes in methodology. Section 4, dealing with what is in essence a small subset of a sensitivity analysis, shows that there are very large bands of uncertainty in final rankings. A fuller sensitivity analysis would vary what factors are included (population, policy and poverty), which

---

[11] Geddes (2012, p. 8) recently stated that 'the conflicting nature of the results they [indices] provide is not helpful for policy-makers'.

variables are chosen to represent them (e.g. logged real GDP *per capita* or GNI in current prices), the scale used to measure them (e.g. the typical index method, a regression or head-count approach), the aggregation method employed (geometric or arithmetic) and other methodological choices (e.g. the base year and sample used). This would clearly give much greater variance of possible judgements than the five measures displayed in Figure 2. Ironically, given the deliberate choice to move away from an econometrically advanced approach, descriptive research may soon be aptly described by Leamer's (1983, p. 37) classic critique of econometric work: 'hardly anyone takes anyone else's data analyses seriously'.
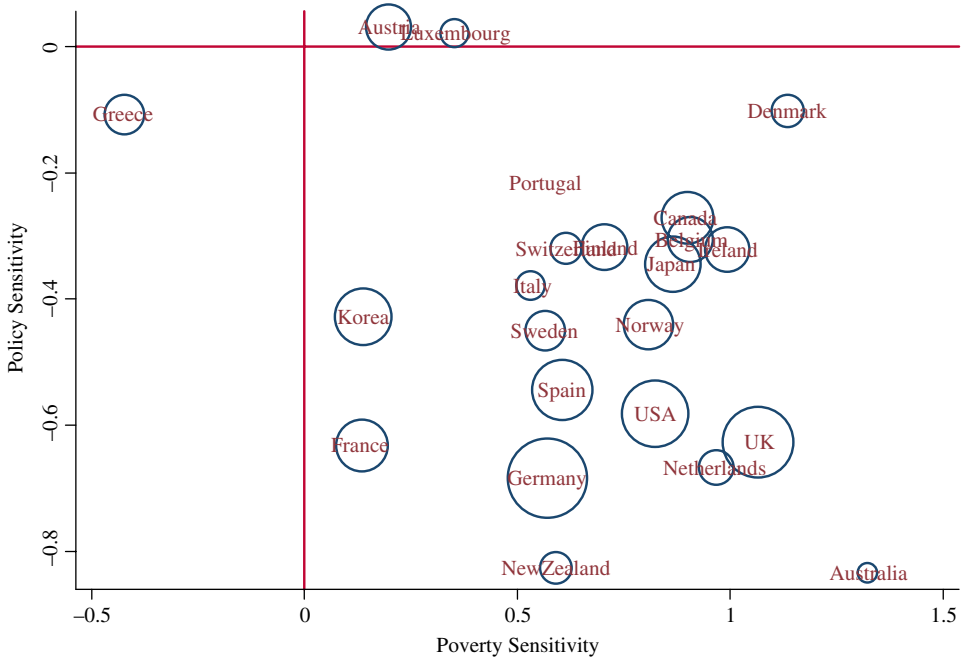
### b. Less Aggregation

An alternative is to present a disaggregated index. This does not avoid all of the problems of the descriptive approach as there is still uncertainty over which factors are relevant and the best way to measure them; both could lead to disparate judgements. However, disaggregation does avoid the pitfall of deciding upon how best to aggregate them, and therefore does not need to decide upon the appropriate MRS. This in itself makes the methodology of measurement less important, as it no longer determines the relevant importance of each factor. Within the descriptive tradition, Nunnenkamp and Thiele (2006) is the only disaggregated presentation of donor allocative performance of which I am aware. This approach is presumably unpopular because of the difficulty in assimilating the information for the reader, due to the lack of a ranking mechanism. For this reason, I follow White and Woestman (1994) who, in similar debate (on general donor performance) on the wisdom and method of aggregation, promoted a graphical alternative to aggregation used by Åshuvud (1986). The graph used was a four axes 'diamond' where the distance from the centre denoted the size of a measure (exactly the same type of graph that is now used by Birdsall and Kharas, 2010). A downside of the graph is that it is not easy to display several donors on the same graph. Donor allocative performance is related to just three factors, making it easier to display graphically.

Figure 3 succinctly displays the sensitivity of 23 donors to poverty, policy and population in absolute terms. Specifically, they are the coefficients taken from three bivariate regressions per donor: logged aid regressed on logged population, logged (PPP) income *per capita* and policy (Worldwide Governance Indicator average). The *x*-axis represents the income coefficient (times by −1 so that it represents poverty not income), the *y*-axis is the policy coefficient and the size of the circle reflects the population coefficient. Thus a donor with high poverty, policy and population sensitivity would have a large circle in the north-east corner of the graph. France has a population coefficient of around 1, whereas Portugal has no circle, as its population coefficient is negative. In Figure 3, only two donors have positive coefficient estimates for policy – Luxembourg and Austria. Greece is a particularly worrying donor here as it gives more aid to richer and less well-governed recipients in absolute terms. The other 20 donors all reside in the south-east quadrant – a clear sign that donors are income sensitive but not policy sensitive (a common finding in the positive literature, see Easterly, 2007; Hout, 2007; Clist, 2011). This point is underlined when examining the scales of the axes. Some donors are remarkably similar – Canada, Belgium, Finland and Japan all overlap. There is also a clear trend for smaller donors to have a greater small country bias, perhaps as they seek to concentrate on recipients for whom they can give relatively substantial aid flows.

While not a perfect solution, this approach maintains the ability for non-specialist readers to quickly assimilate the information. The disaggregated form also makes clear the trade-offs that donors have made: some may be more poverty sensitive, whereas others are more popula-

FIGURE 3
Donor Performance using Bivariate Regression Coefficients, 2009



Source: Author's calculations, using 2009 data. Coefficients taken from 3 bivariate regressions.

tion sensitive. By aggregating, the researcher imposes a marginal rate of substitution, and the donors are ranked according to that MRS. The graphical approach makes clear the MRS that each donor implicitly uses. Roodman (2011b, p. 483) admits that 'aggregation hides at least as much as it reveals'. While the approach taken here is not amenable to an index with more factors, the graphical presentation reveals more than a simple index in the case of donor allocative performance. It is said that a picture is worth a thousand words: Figure 3 is worth some 69 bivariate regressions.

### c. Abandon Indices

Given the evidence presented earlier, it is worth fundamentally questioning the value of the descriptive approach. Final rankings are fragile to minimal changes in methodology, and popular measures have a very low degree of consensus. Is a graphical option (or other disaggregated approach) a credible way forward, given that it does not solve all of the problems associated with current rankings? There are two aspects of the descriptive approach that make it (potentially) worthwhile: one technical and the other pragmatic. First, I argue, like Easterly and Williamson (2011), that the advantage comes from using absolute judgements to complement explanatory approaches. Descriptive analysis is unique in that it can answer questions regarding absolute poverty, population and policy sensitivity, where more econometrically advanced methods are only able to answer such questions conditional on other factors.

Second, while rankings (and graphical presentations) based on simple measures may not be a first-best solution, they have become popular because of an understandable desire to judge donor performance. Furthermore, it is clear that the descriptive approach can be more simply understood by non-specialists, in a way that is not the case with more econometrically advanced work. This advantage threatens to be undermined by fragile and contradictory judgements, but the potential advantage should not be dismissed.

The reader must judge whether these two reasons are enough to counteract the weaknesses of the approach. Pragmatically however, it appears likely that descriptive approaches will continue to be used regardless of their fragility. It is in this context that I recommend graphical presentations that avoid aggregation but maintain a simple presentation of donor performance: not as a first-best solution but as an improvement on fragile rankings.

## 6. CONCLUSION

The aim of descriptive measures is laudable: to improve aid effectiveness by encouraging good allocative practice. Because the imagined mechanism for changing allocative behaviour includes increased public awareness of the issues, simplicity in measurement and presentation has been valued. The latter has been fruitful, as rankings have successfully increased public discussion and involvement in the issues of donor allocative behaviour. However, it appears naive to claim that in favouring simpler methods, the descriptive literature has minimised methodological uncertainty. I have shown that even minimal changes to methodology result in different rankings, and often in very different rankings. Changes such as whether to include or exclude a single recipient, the variable used and the base year lead to non-trivial changes. The choice of factors, measurement approach and aggregation technique would logically lead to even larger differences in final rankings. The aggregate effect of the inherent whimsy in popular descriptive measures is that of substantial sensitivity to methodology. Similarly, competing measures, implemented using the canonical model, give contradictory judgements. The range of possible opinions in a full sensitivity analysis of selectivity would surely be even larger than the actual range of opinions discussed here. The lack of clear theoretical or technical dominance of one approach over another means a plurality of opinions are potentially valid. This sensitivity to small technical details and disagreement in final rankings threatens to undermine increased public awareness, as methodological ambiguity belies the false certainty of unambiguous rankings. In future, donors would not need to claim that a poor judgement of their performance was a 'certain value judgement'; it could merely point to a competing measure that praised their performance. Indeed, Kihara (2012, p. 3–4) recently defended Japanese aid using such an argument: 'some empirical studies, including studies presented here, seem to contradict the findings of the CGD. A number of recent studies have ranked aid donors by various indicators of their aid-giving and in some of these Japanese aid has been ranked toward the upper end of donor countries'. The problems with indices notwithstanding, their advantages mean demand for them is likely to continue, and with this in mind, a graphical approach is recommended as a way of avoiding some of the largest problems. Much, but certainly not all, of the disagreement in final rankings can be avoided if distinct elements of an index are left in disaggregated form. This approach has not been embraced as disaggregation has typically meant poor presentation, which would undermine the intermediate goal of public communication and engagement. Section 5 proposes a graphical solution to the resulting impasse, combining disaggregation with a presentation that is easy to understand. The graph displays the poverty, policy and population sensitivity of 23 donors in absolute terms

(although the graph could equally display conditional results). In doing so, the marginal rate of substitution is not imposed on the rankings, but can be inferred by the reader. In fact, the graph does not even impose decision of the appropriate factors, as a reader could choose to ignore one of the three dimensions. I argue that the descriptive approach is a useful complement to explanatory research. In that vein, recent findings of low policy sensitivity and high income sensitivity in explanatory research (Easterly, 2007; Hout, 2007; Clist, 2011) is echoed in Figure 3. This enables researchers to have more confidence in the result as this is found in both conditional and absolute terms using explanatory and descriptive approaches. Measuring absolute sensitivity (i.e. not controlling for other factors) also makes clear the inherent dichotomy that donors face (see Figure 1): richer countries tend to be better governed. As such the advice of Collier and Dollar (2002) to be both income and policy sensitive is easier to give than to follow; the graphical method confirms that donors generally choose the former. The method makes clear the difficult choice that donors face, but also highlights the donors that fare poorly. By measuring donor performance better, it is hoped that unprofitable debate over the correct index is avoided, and the focus remains on increasing donor performance by measuring it.

## REFERENCES

Alesina, A. and D. Dollar (2000), 'Who Gives Foreign Aid to Whom and Why?', *Journal of Economic Growth*, **5**, 1, 33–63.

Anderson, E. and P. Clist (2011), 'Measures for Measures: Evaluating Judgements of Donor Allocative Performance', UEA Working Paper 35 (Norwich: University of East Anglia).

Åshuvud, J. (1986), 'Statistical Review of Swedish Development Aid', in P. Frühling (ed.), *Swedish Development Aid in Perspective* (Stockholm: Almqvist and Wiksell), 283–312.

Berthélemy, J. C. and A. Tichit (2004), 'Bilateral Donors' Aid Allocation Decisions–a Three-dimensional Panel Analysis', *International Review of Economics & Finance*, **13**, 3, 253–74.

Birdsall, N. (2011), 'Comment on Multi-dimensional Indices', *Journal of Economic Inequality*, **9**, 3, 489–91.

Birdsall, N. and H. Kharas (2010), Quality of Official Development Assistance Assessment (Washington, DC: Center for Global Development).

Clist, P. (2011), '25 Years of Aid Allocation Practice: Whither Selectivity?', *World Development*, **39**, 10, 1724–34.

Collier, P. and D. Dollar (2002), 'Aid Allocation and Poverty Reduction', *European Economic Review*, **46**, 8, 1475–500.

Easterly, W. (2007), 'Are Aid Agencies Improving?', *Economic Policy*, **22**, 52, 633–78.

Easterly, W. and C. Williamson (2011), 'Rhetoric versus Reality: The Best and Worst of Aid Agency Practices', *World Development*, **39**, 11, 1930–49.

Geddes, M. (2012), 'Where do European Institutions Rank on Donor Quality?', ODI Background Note (London: Overseas Development Unit).

Hout, W. (2007), *The Politics of Aid Selectivity: Good Governance Criteria in World Bank, US and Dutch Development Assistance*. (London: Taylor & Francis).

Høyland, B., K. Moene and F. Willumsen (2012), 'The Tyranny of International Index Rankings', *Journal of Development Economics*, **97**, 1, 1–14.

Kihara, T. (2012), 'Effective Development Aid: Selectivity, Proliferation and Fragmentation, and the Growth Impact of Development Assistance', ADBI Working Paper Series, 342 (Tokyo: Asian Development Bank Institute).

Kilby, C. and A. Dreher (2010), 'The Impact of Aid on Growth Revisited: Do Donor Motives Matter?', *Economics Letters*, **107**, 3, 338–40.

Klugman, J., F. Rodríguez and H. J. Choi (2011), 'The HDI 2010: New Controversies, Old Critiques', *Journal of Economic Inequality*, **9**, 2, 249–88.

Knack, S., F. H. Rogers and N. Eubank (2011), 'Aid Quality and Donor Rankings', *World Development*, **39**, 11, 1907–17.

Leamer, E. E. (1983), 'Let's Take the Con Out of Econometrics', *The American Economic Review*, **73**, 1, 31–43.

Llavador, H. G. and J. E. Roemer (2001), 'An Equal-opportunity Approach to the Allocation of International Aid', *Journal of Development Economics*, **64**, 1, 147–71.

McGillivray, M. (1989), 'The Allocation of Aid Among Developing Countries: A Multi-donor Analysis Using a *per capita* Aid Index', *World Development*, **17**, 4, 561–68.

McGillivray, M. (1991), 'The Human Development Index: Yet Another Redundant Composite Development Indicator?', *World Development*, **19**, 10, 1461–68.

McGillivray, M. (1992), 'Reply', *World Development*, **20**, 11, 1699–702.

MOFA (2006), 'Critical Comments on the Ranking of Developed Countries Made by CGD, A US Non-governmental Think Tank', Available at: http://www.mofa.go.jp/policy/oda/other/index0609.html (accessed 1 October 2011).

Nunnenkamp, P. and R. Thiele (2006), 'Targeting Aid to the Needy and Deserving: Nothing But Promises?', *The World Economy*, **29**, 9, 1177–201.

Ravallion, M. (2011), 'On Multidimensional Indices of Poverty', *Journal of Economic Inequality*, **9**, 2, 235–48.

Ravallion, M. (2012), 'Mashup Indices of Development', *World Bank Research Observer*, **27**, 1, 1–32.

Roodman, D. (2004), 'An Index of Donor Performance', Working Paper 42 (Washington, DC: Center for Global Development).

Roodman, D. (2006), *The Commitment to Development Index*: *2006 Edition* (Washington, DC: Center for Global Development).

Roodman, D. (2007), 'The Anarchy of Numbers: Aid, Development, and Cross-country Empirics', *World Bank Economic Review*, **21**, 2, 255–77.

Roodman, D. (2011a), 'An Index of Donor Performance', Working Paper 67 (Washington, DC: Center for Global Development).

Roodman, D. (2011b), 'Composite Indices', *Journal of Economic Inequality*, **9**, 3, 483–84.

White, H. (1992), 'The Allocation of Aid Among Developing Countries: A Comment on McGillivray's Performance Index', *World Development*, **20**, 11, 1697–98.

White, H. and M. McGillivray (1995), 'How Well is Aid Allocated? Descriptive Measures of Aid Allocation: A Survey of Methodology and Results', *Development and Change*, **26**, 1, 163–83.

White, H. and L. Woestman (1994), 'The Quality of Aid: Measuring Trends in Donor Performance', *Development and Change*, **25**, 3, 527–54.

## APPENDIX

This appendix includes the rankings and rescaled raw scores used to make Figure 2. The methods used are those of the original measures, with data and Stata do files available to download from  https://sites.google.com/site/paulclist/data. The rankings may differ from elsewhere in the paper as I restrict the measures in the table to use the same sample.

TABLE A1
Raw Scores and Rankings for Bilateral Donors, 2009

| Donor | Rescaled Raw Performances | | | | | Rankings | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | EW | API | MPI | RPI | KRE | EW | API | MPI | RPI | KRE |
| Australia | 82.1 | 21.4 | 92.4 | 71.8 | 4 | 4 | 13 | 8 | 4 | 22 |
| Austria | 0 | 28.9 | 94.9 | 46.7 | 62.6 | 22 | 8 | 4 | 11 | 15 |
| Belgium | 28.4 | 12.5 | 94.2 | 28.3 | 86.5 | 16 | 19 | 5 | 18 | 5 |
| Canada | 53.7 | 21.8 | 81.4 | 59.8 | 80 | 9 | 12 | 16 | 6 | 9 |
| Denmark | 69.5 | 20.2 | 84.8 | 94.3 | 69.9 | 7 | 15 | 13 | 2 | 12 |
| Finland | 71.6 | 29.8 | 92.6 | 57.2 | 61.4 | 6 | 7 | 7 | 10 | 16 |
| France | 43.2 | 25.5 | 0 | 22.3 | 48.2 | 12 | 9 | 23 | 20 | 18 |
| Germany | 46.3 | 93.1 | 89.6 | 37.6 | 98 | 11 | 2 | 9.5 | 15 | 2 |
| Greece | 7.4 | 3.1 | 76.4 | 0 | 34.3 | 20 | 22 | 19 | 23 | 20 |
| Ireland | 64.2 | 31.9 | 97.4 | 80.3 | 87.7 | 8 | 4 | 3 | 3 | 4 |
| Italy | 2.1 | 19.2 | 88.5 | 9.7 | 35.7 | 21 | 17 | 12 | 22 | 19 |
| Japan | 74.7 | 89.5 | 50.5 | 59.2 | 86.1 | 5 | 3 | 20 | 8 | 6 |
| Korea | – | 30.9 | 81.2 | 44 | 71.7 | – | 6 | 17 | 13 | 11 |
| Luxembourg | 89.5 | 12.3 | 84.1 | 69.2 | 67.3 | 2 | 20 | 15 | 5 | 14 |
| Netherlands | 33.7 | 22 | 11.1 | 35.9 | 69.7 | 15 | 11 | 22 | 17 | 13 |
| New Zealand | 100 | 5.3 | 100 | 18.2 | 9.3 | 1 | 21 | 1 | 21 | 21 |
| Norway | 38.9 | 30.9 | 77.1 | 36.9 | 85.9 | 13 | 5 | 18 | 16 | 7 |
| Portugal | 35.8 | 0 | 99.8 | 100 | 0 | 14 | 23 | 2 | 1 | 23 |
| Spain | 51.6 | 15.2 | 89.6 | 43.2 | 81.6 | 10 | 18 | 9.5 | 14 | 8 |
| Sweden | 21.1 | 21.3 | 84.2 | 45.9 | 74.1 | 18 | 14 | 14 | 12 | 10 |
| Switzerland | 23.2 | 20 | 93.1 | 58.3 | 49.7 | 17 | 16 | 6 | 9 | 17 |
| UK | 84.2 | 100 | 89.2 | 59.4 | 94.5 | 3 | 1 | 11 | 7 | 3 |
| USA | 12.6 | 22.9 | 23.7 | 24.2 | 100 | 19 | 10 | 21 | 19 | 1 |