Costing a Data Revolution

Gabriel Demombynes and Justin Sandefur

Abstract

The lack of reliable development statistics for many poor countries has led the U.N. to call for a "data revolution" (United Nations, 2013). One fairly narrow but widespread interpretation of this revolution is for international aid donors to fund a coordinated wave of household surveys across the developing world, tracking progress on a new round of post-2015 Sustainable Development Goals. We use data from the International Household Survey Network (IHSN) to show (i) the supply of household surveys has accelerated dramatically over the past 30 years and that (ii) demand for survey data appears to be higher in democracies and more aid-dependent countries. We also show that given existing international survey programs, the cost to international aid donors of filling remaining survey gaps is manageable--on the order of $300 million per year. We argue that any aid-financed expansion of household surveys should be complemented with (a) increased access to data through open data protocols, and (b) simultaneous support for the broader statistical system, including routine administrative data systems.

JEL Codes: C82, F35, O10

Keywords: household surveys, national statistics, open data, aid, Sustainable Development goals.

Center for Global Development
www.cgdev.org

Working Paper 383
October 2014

Costing a Data Revolution

Gabriel Demombynes
World Bank and NYU Financial Access Initiative

Justin Sandefur
Center for Global Development

We are grateful to Olivier Dupriez, Claire Melamed, and Amanda Glassman for comments on an earlier draft, and to Olivier for help in understanding the IHSN data. All remaining errors are ours. The views expressed in this paper are the authors' alone. They do not necessarily reflect the views of the World Bank or its Executive Directors, the Center for Global Development, its board, or its funders.

-Senior Economist, World Bank, gdemombynes@worldbank.org.

-Research Fellow, Center for Global Development, jsandefur@cgdev.org.

CGD is grateful for contributions from the Omidyar Network, the UK Department for International Development, and the William and Flora Hewlett Foundation in support of this work.

Gabriel Demombynes and Justin Sandefur. 2014. "Costing a Data Revolution." CGD Working Paper 383. Washington, DC: Center for Global Development. http://www.cgdev.org/publication/costing-data-revolution-working-paper-383

Center for Global Development
2055 L Street, NW
Fifth Floor
Washington, DC 20036

202.416.4000
(f) 202.416.4050

www.cgdev.org

The Center for Global Development is an independent, nonprofit policy research organization dedicated to reducing global poverty and inequality and to making globalization work for the poor. Use and dissemination of this Working Paper is encouraged; however, reproduced copies may not be used for commercial purposes. Further usage is permitted under the terms of the Creative Commons License.

The views expressed in CGD Working Papers are those of the authors and should not be attributed to the board of directors or funders of the Center for Global Development.

1 Introduction

As 2015 approaches, a debate has emerged on what will follow the Millennium Development Goals. The UN High Level Panel on new goals has called for a "data revolution" to monitor development progress (United Nations, 2013). There is no clear consensus on what a data revolution means. The lead author of the UN High Level Panel suggests the panel did not have any particular model in mind (Kharas, 2014). Roodman (2014) notes four interpretations of the "data revolution" in policy conversations: (i) a technology revolution, (ii) open data, (iii) capacity building in national statistics agencies, and (iv) a big survey push.

The Copenhagen Consensus analysis in Jerven (2014) focuses on the fourth proposal: a new, expanded, and globally coordinated wave of household surveys to measure progress on each new post-2015 UN development indicator on a regular basis across the developing world. Similar proposals for a big survey push have been advanced by Alkire and Samman (2014) among others. In this paper, we take an empirical look at three questions related to this vision of the data revolution.

First, has the previous push for household surveys in poor countries produced results? To put the data revolution debate in context, we draw on data from the International Household Survey Network (IHSN) to examine the pace of survey production over time and across countries. We show that both survey production and availability have increased dramatically over time and find that poorer countries now conduct *more* household surveys than middle-income countries, and are *more* likely to put them in the public domain.

Second, what types of users demand what types of data? Over (2014) proposes four categories of data consumers in developing countries: citizens, government, foreign investors, and international donors. Sandefur and Glassman (2014) argue that nationally representative surveys are often designed to suit the demands of international donors making cross-country comparisons, rather than governments and citizens doing sub-national analysis. Nevertheless, both survey data in general and open data in particular may be driven in part by citizen demand. Regressions using the IHSN data provide findings compatible with both hypotheses: both survey production and public data dissemination increase with a country's Polity IV democracy score, and data openness rises significantly as countries become more aid dependent, suggesting a large role for foreign donors in the demand for data access.

Third, how much would it cost to close the remaining gaps in household survey production? We take the calculations made for the Copenhagen Consensus by Jerven (2014) of the

total costs of a basic data package across all developing countries as a reasonable benchmark. However, we argue that the resulting figure from that calculation gives an exaggerated sense of the international funds needed to close existing gaps, because middle-income countries can finance surveys with domestic resources. Focusing on countries below $2,000 per capita GDP in PPP dollars yields a total cost to international donors of closing all remaining survey gaps of less than $300 million per annum – a fairly small share of global aid budgets.

Note that we do not address the emerging role of new technologies in data capture, such as the use of cell phone data, remote sensing via satellite, or data exhaust from online transactions. Our view is that these new approaches will complement rather than substitute for traditional surveys as part of an integrated national statistics system.

The rest of the paper is organized as follows. The next section presents our analysis of the IHSN database, documenting trends in survey production and cross-country correlates of both data production and open data. Section 3 presents our cost calculations to raise all countries to a minimum benchmark of survey coverage. The remaining sections reflect on the policy implications of these calculations. Section 4 highlights the importance of open data policies, while Section 5 steps back to consider the broader goals of a "data revolution." While survey data is critical for many purposes, including monitoring national progress on international goals, delivering basic services to achieve those goals requires additional focus on the broader tasks of statistical capacity building, including routine administrative statistics and other sub-national data systems.

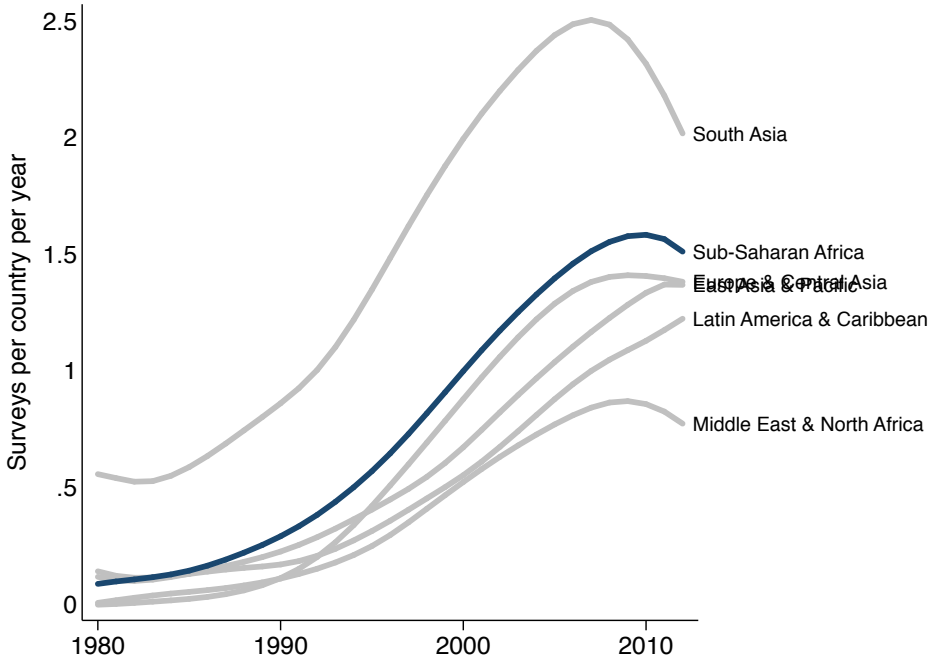
2 Trends and Correlates of Data Production and Openness

To better understand the state of both data production and openness, we examine data from the International Household Survey Network (IHSN) database, which is the most comprehensive collection of information regarding surveys and censuses from low- and middle-income countries. It includes data on survey type and country as well as whether the microdata is in the "open", meaning accessible on-line.[1] We match the IHSN data with a series of

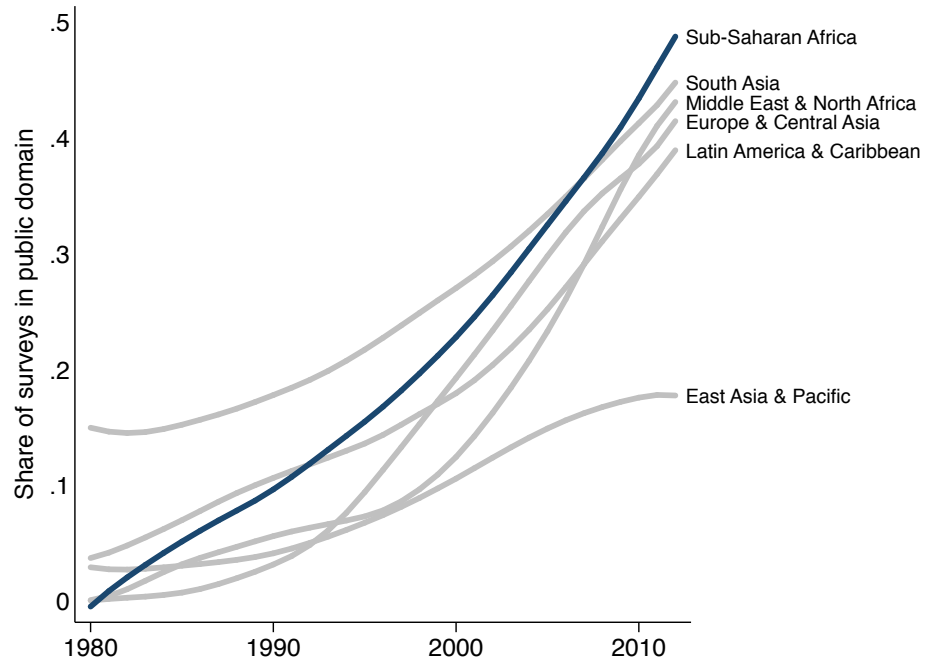
[1]The term "open data" is sometimes used specifically to refer to data that is made available with no restrictions whatsoever and not in a proprietary format. In this paper we use the term more broadly to refer to cases of microdata that is made available for public use at no cost. As we use the term, it includes the many cases where modest conditions are attached to the use of the data, e.g. users may not attempt to

Figure 1: Trends in survey data collection and dissemination

(a) Number of surveys



(b) Share of survey data open to public



country-level characteristics drawn from various sources. We begin by examining simple trends over time in the prevalence of survey data collection and publication of open data, and then examine country-level correlates of these outcomes in a regression framework.

The IHSN data allow us to address two main questions. First, has the previous push for household surveys in poor countries produced results? A number of recent analyses have decried the lack of reliable development statistics for poor countries, particularly in sub-Saharan Africa (Devarajan, 2013; Jerven, 2013; Sandefur and Glassman, 2014). But is the problem too few household surveys, or some other combination of factors, such as unreliable survey data, deficiencies in administrative data systems, or flawed data analysis and reporting?

The pace of survey data collection has accelerated rapidly across all regions, as shown in Figure 1a. Using the World Bank's regional classifications, South Asia reports the highest rate of data collection, with about 2 surveys per annum as of 2010. Rather than trailing the world, sub-Saharan Africa ranks second among global regions as of 2010, at about 1.5 surveys per annum. [2]

The increasing trend toward making data open is also ubiquitous across regions. Almost all regions, with a minor exception for South Asia, show virtually zero open data as of the beginning of the series in 1980. By 2011 most regions are clustered at around 40% of surveys in the public domain, with sub-Saharan Africa leading the world at roughly 50%. East Asia and the Pacific is an outlier, remaining below 20% even in the most recent years.

The regressions in Table 1 paint a consistent picture. Poorer countries produce significantly more household surveys per annum, as seen in column 1, which uses the full sample of 180 countries. A one log point increase in per capita GDP in PPP dollars is associated with 0.23 fewer surveys per annum. Conditional on running a survey, poorer countries are

identify respondents or sell the data, as well as cases where the data is distributed in a proprietary format (such as Stata or SPSS format). For the analysis of the IHSN database we consider a dataset "open" if is accessible on-line, meaning that the microdata can be obtained on-line free of charge and without severe or unknown restrictions. Surveys that countries share in their catalogs under "licensed access" are not included. Data are available on the IHSN website at http://catalog.ihsn.org. The figures we report are based on data as of April 8, 2014.

[2]The apparent drop-off in the number of surveys in recent years reflects the fact that the microdata from many recent surveys has not yet been yet released, and thus those surveys do not appear in the IHSN database. Likewise, part of the reason that the number of surveys shown in the 1980s is low is that the IHSN database has prioritized more recent surveys, and not all earlier surveys have been documented and entered into the database. Adding the missing surveys, however, would not substantially change the picture presented here.

also *more* rather than less likely to publish open data. One log point in per capita GDP is associated with a roughly 5% decrease in the proportion of existing survey that are available publicly in some form.

In sum, the data show that poorer developing countries already collect more household survey data than their middle-income counterparts, and are more likely to put their survey data in the public domain. Nevertheless, worldwide roughly half of surveys and censuses are still not publicly available. Openness may not be a major concern for the particular issue of constructing national-level indicators for the purpose of monitoring international goals. But when microdata is only available within a narrow circle, opportunities are limited to use the data for analysis to inform national policy as well as draw cross-country policy lessons.

Now we turn to the second question we attempt to answer with the IHSN data: who demands data – both survey data in general, and open data in particular? Open access to nationally representative survey data may be driven by citizen demand. In addition open access to survey data may also result from demand from international aid donors.

The data offer some support for both hypotheses – with the obvious caveat that we are looking at mere correlations with a fairly limited set of controls and no basis for causal inference here. Countries that receive more foreign aid show no significant tendency to conduct more household surveys, but they are more likely to publish open data by a modest but statistically significant margin. A one standard deviation increase in the log aid share of the government budget is associated with an 8% increase in the share of survey microdata that is released. The role of aid in data access is also clear when looking at the composition of surveys. The relatively high percentage of surveys in the poorest countries (and specifically in Africa) which are in the public domain is in part a consequence of the fact that a substantial number are internationally-sponsored surveys like the Demographic and Health Surveys (DHS) and Multiple Indicator Cluster Surveys (MICS), which are generally publicly available.

Turning to the role of domestic accountability, we find that for aid, more democratic countries are also more likely to publish their data. A one standard deviation increase in the Polity IV democracy score is associated with 0.33 more surveys per annum and a 4% increase in the fraction of surveys that are made available publicly. Both results are significant at the 10% level. One immediate doubt about this association is whether it captures the role of democratic accountability, or simply better quality institutions that generally provide better government services. To address this concern, we control for the Worldwide Governance

Table 1: Cross-country regressions using the IHSN catalog

	Surveys per year			Share of surveys open		
	(1)	(2)	(3)	(4)	(5)	(6)
Log per capita GDP, PPP	-0.234***	-0.305	-0.375	-0.0481***	-0.0258	0.0299
	(0.0431)	(0.190)	(0.276)	(0.00617)	(0.0199)	(0.0357)
Log population	0.194***	0.567***	0.676***	0.0277***	0.00742	0.0293*
	(0.0372)	(0.130)	(0.137)	(0.00397)	(0.0135)	(0.0157)
Log aid share of gov. budget			0.178			0.0427**
			(0.124)			(0.0164)
Polity IV democracy index			0.330*			0.0407*
			(0.177)			(0.0228)
WGI gov. effectiveness index			1.015**			0.0446
			(0.414)			(0.0351)
Observations	4179	730	730	4179	730	730
Countries	180	94	94	180	94	94
R-squared	0.21	0.21	0.28	0.19	0.14	0.16

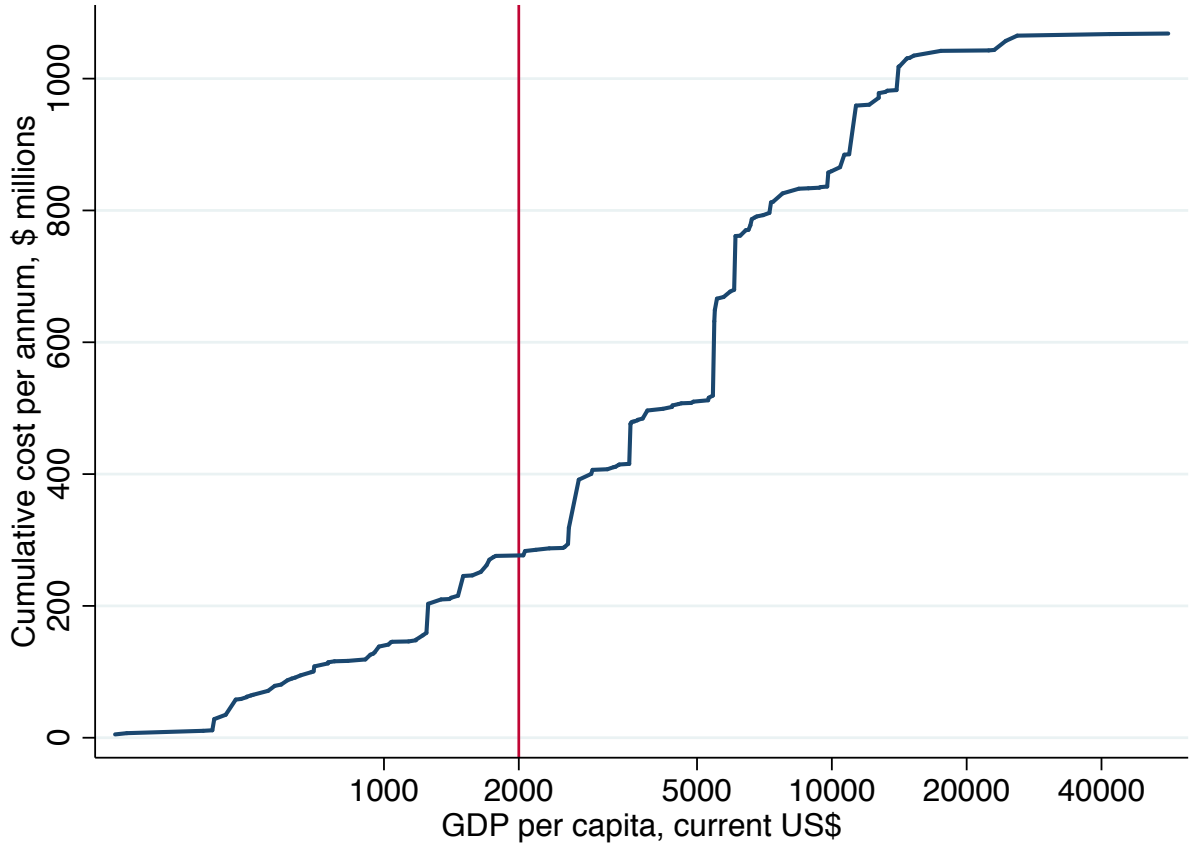
Note: The data set consists of one observation per country per year, from 1990 to 2014. In the first three columns, the dependent variable is the total number of surveys conducted per annum. In the last three columns, the dependent variable is the share of these surveys for which data is freely accessible online. All equations include year fixed effects. Standard errors are clustered at the country level. Asterisks (*,** ,***) denote coefficients that are significantly different from zero at the 10, 5 and 1% levels, respectively.

Indicators (WGI) measure of government effectiveness. The WGI indicator is a significant, positive correlate of survey data production, but is insignificant in the open data regression.

3 The Costs of Basic Data Production

How much money is needed to support a program of minimal statistics in the developing world? The Copenhagen Consensus analysis conducted by Jerven (2014) estimates that the total cost for producing the data for the post-2015 goals (often referred to as Sustainable Development goals, henceforth SDGs) over 1990-2015 would have been $27 billion, or just over $1 billion per year over the SDG period. The $1 billion figure is based on a package of data collection that includes a population census every ten years, a Demographic and Health Surveys (DHS) every 5 years, a Living Standards Measurement Study every five years, and an annual Core Welfare Indicator Questionnaire. Without necessarily endorsing exactly this

Figure 2: Annual cost of full household survey schedule in all countries below a given GDP threshold



data package, we take it as a reasonable, approximate guideline for minimal socioeconomic data needs.

What this calculation neglects is that many of the countries included are wealthy enough to fully fund their own statistical apparatuses, and many already do. Surveys and censuses in Kuwait, South Korea, and Chile for example, are included in the $1 billion figure. Using unit-cost data from Jerven (2014), Figure 2 shows a plot of GDP per capita of countries ranked from low to high versus cumulative costs. Thus for any particular point, the value on the vertical axis is the annual cost of funding the minimal SDG program for all countries with GDP per capita up to and including that point.

Recognizing the international public good value of socioeconomic data and limited funding out of own-budgets in the poorest countries, we would expect international development assistance to fund a substantial fraction of statistical costs for those countries below some cutoff level, and only a small share costs above that level. We suggest that this point is somewhere within the range of US$2000-$5000. The total costs of the data package are $275 million per year for all countries with GDP per capita under $2000 and $510 million for all those under $5000 GDP per capita. The bulk of countries with GDP under $2000 (36 out of 52) are in sub-Saharan Africa. The total annual data cost for all countries in sub-Saharan Africa is $276 million. Based on these figures, we suggest that the total amount of international donor assistance needed to support this basic survey program is on the order of $300 million per year.

It is also worth noting that current international donor assistance flows already provide a substantial portion of these funds. USAID funds the DHS, concessional International Development Assistance (IDA) loans partially fund LSMS surveys in many low-income countries, and UNFPA and other agencies provide financial supports to censuses.

Finally, we highlight that our calculations are for a minimum statistical package in the developing world. It is incorrect to view $300 million as an estimate of the *additional* donor assistance needed just to monitor the post-2015 version of the Millennium Development Goals. That is because at least some elements of the data package underlying the calculation will need to be carried out for other purposes. With or without a new set of goals, countries will continue to carry out censuses and surveys every few years. In fact the additional cost of monitoring a new set of goals, beyond what countries need for their own purposes, is quite low.

We draw two conclusions from this brief analysis. First, while $1 billion per year is a reasonable order-of-magnitude calculation of the costs of producing the specified data package, the costs of international funding support to produce the specified data package are much less—very roughly $300 million. Second, $300 million is a total cost figure, not the marginal increase in aid required. Starting from current funding levels, the additional funding needed *above* current levels of support to bring all developing countries up to a basic level of socioeconomic data production is likely to be considerably less than $300 million per year.

4 The Importance of Data Openness

While additional international funding for data production is needed, we argue that greater efforts to ensure data openness are of equal importance. In many cases, microdata from a household survey or census is collected and then used to produce a single report, remaining afterwards in the electronic equivalent of a dusty and forgotten cabinet drawer. For administrative data like that collected by health and education ministries, the situation is typically even worse: great effort is expended to collect detailed school and health clinic data, and the data is never used for anything beyond producing a few aggregate summary statistics.

One reason that data is hidden away is that data producers are often embarrassed by the quality of the underlying data and unwilling to have someone sniffing around pointing out problems. A second reason is that data is power, but not in a good way. Organizations keep a tight grip on their data because it is a thing of value. As long as they hold exclusive access, they have the possibility of receiving contracts for analyzing the data or outright selling the data.

One of many examples is the 2005-06 Kenya Integrated Household Budget Survey (KIHBS), the countrys most recent multi-purpose consumption survey conducted. This survey should be a keystone reference for understanding poverty, agriculture, employment and many other issues. Unfortunately, although in principle the data is available on request from the Kenya National Bureau of Statistics, in practice it has been made available to only a very small circle of researchers under the proviso that it not be shared more widely.

With few exceptions data collection in developing countries has been paid for with public money, either by the country citizens paying taxes to their governments or by taxpayers

abroad who fund bilateral and international organizations that support data collection. It is those citizens who are the rightful owners of that data. As a broad principle, publicly funded data should be freely available to the public within 1-2 years of collection.

Of course this principle should be subject to some conditions. Individual identifying information should be stripped for datasets, and adequate time should be allowed for the researcher or data producer to process and take a "first cut" at the data.

There are already a number of laudable data access models, such as the Afrobarometer, the Demographic and Health Surveys, the INDEPTH data repository, the Living Standards Measurement Study Integrated Survey on Agriculture, and the International Integrated Public Use Microdata project. Making these models the rule rather than the exception will require governments and organizations that fund data collection to do two things: 1) make open data access the norm for funding agreements, and 2) ensure that data dissemination is funded from the start.

5 The Right Goals for International Statistics

A narrow focus on data for post-2015 international goal monitoring could potentially distort the broader push for improving statistics in developing countries (Data for African Development Working Group, 2014). The main value of data is not for monitoring international goals but to generate knowledge for policy and economic decision-making in each country.

Returning to the discussion in the introduction, various categories of data users – citizens, governments, foreign investors, and aid donors – require different types of statistics. In the education sector, for instance, aid donors might monitor national averages for enrollment or test score performance on some internationally comparable metrics such as those proposed for the SDGs. Meanwhile, Ministry of Education officials may wish to disaggregate these data to the level of districts, both to evaluate the performance of district officials and to allocate resources such as additional teachers and textbooks. At an even lower level, parents may desire highly disaggregated information on the performance of the schools within travel distance from their homes.

The education example reinforces the importance of open data access discussed above. It also highlights the need to go beyond surveys that produce a single, nationally representative

statistic once every several years. While this may suffice for international donors, many domestic users require much more disaggregated data at much higher frequency.

These different needs present a trade-off in funding statistics. Highly aggregated, low-frequency survey data are often more closely scrutinized and thus more reliable than disaggregated, high-frequency administrative data sources. Unfortunately, on key indicators in education and health, these two sources often disagree, as Sandefur and Glassman (2014) show for a panel of African countries. For instance, in Kenya primary enrollment rates in survey data remained stagnant after the abolition of user fees in 2003, while administrative data – potentially driven by incentives to over-report enrollment and collect additional per pupil funding – showed rapid increases. Rather than undermining the need for household surveys, these discrepancies highlight the need for better integration of administrative and survey data sources to improve the accuracy of the former and the relevance of the latter.

Finally, the push to expand household surveys to monitor the SDGs could inadvertently lead to a take-over of data collection responsibilities by foreign donors at the expense of national statistics offices. We would argue that defining and measuring development indicators must remain the responsibility of country governments. The effort to generate data for international goals should not displace the focus from building the capacity of national statistical agencies.

Household surveys – whether designed and administered by countries or handed down by aid donors – are an important tool for monitoring progress on international development goals. But actually achieving those goals requires governments to deliver basic services like health, education, water, power, and policing to citizens in an efficient and equitable manner. Doing so requires a reliable, integrated national statistics system that provides policymakers the information they need in the frequency and level of aggregation required. In short, household survey data will be useful for monitoring the new SDGs; actually achieving them will require greater focus on other types of data, including administrative systems. The focus on monitoring should not detract from this broader goal.

References

Alkire, S., and E. Samman (2014): "Mobilising the Household Data Required to Progress toward the SDGs," OPHI Working Paper.

Data for African Development Working Group (2014): "Delivering on the Data Revolution in Sub-Saharan Africa: Final Report," Center for Global Development and the African Population and Health Research Council.

Devarajan, S. (2013): "Africa's Statistical Tragedy," Review of Income and Wealth.

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

——— (2014): "The cost and benefit of data needed to monitor post-2015 MDG," Draft paper for the Copenhagen Consensus.

Kharas, H. (2014): "A Data Revolution for the post-2015 Agenda?," World Bank, Future Development blog: http://blogs.worldbank.org/futuredevelopment/data-revolution-post-2015-agenda, Published online, October 1, 2013.

Over, M. (2014): "Using 'Value of Information' Concepts to Prioritize the Data Revolution," Center for Global Development: http://www.cgdev.org/blog/using-value-information-concepts-prioritize-data-revolution, Published online, March 28, 2014.

Roodman, D. (2014): "Interpreting the Data Revolution: Proceed with Caution (Part 1)," Post2015.org: http://post2015.org/2014/04/03/interpreting-the-data-revolution-proceed-with-caution-part-1/, Published online, April 3, 2014.

Sandefur, J., and A. Glassman (2014): "The Political Economy of Bad Data: Evidence from African Survey and Administrative Statistics," Journal of Development Studies, forthcoming.

United Nations (2013): "Communiqué," Meeting of the HighLevel Panel of Eminent Persons on the Post2015 Development Agenda in Bali, Indonesia, 27 March 2013.