RTI Press

Policy Brief

June 2015

A Relevant Data Revolution for Development

Luis Crouch

In 2013, the United Nations issued the highly anticipated *Report of the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda.*[1] Based on its pithy and under-specified—but powerful—call for a data revolution for development, many institutions have started taking action in this area.* The institutions are producing their own strategies and coordinating with other agencies on what this data revolution might mean. For instance, the World Bank issued a landmark report that outlines best practices, needs, and other relevant information.[2] In addition, special committees and inter-institutional working groups (e.g., PARIS21) have been established, and think tanks (e.g., the Center for Global Development) have held workshops and issued publications[3] on the topic.†

Key Findings

While big data offers potential benefits in many realms, applying it to social and economic development problems, as the underpinning of a data revolution for development, could be misleading. A more relevant approach to a data revolution for development would consist of the following four key actions:

1. Improve "little data" and the systems that already house it, and add citizen input and citizen feedback data by using appropriate media.
2. Better integrate, curate, and classify existing data (including "little data") and new data sources.
3. Add value to data: Analyze it to identify over- and underperforming service units, encourage citizen interaction with services, and promote both accountability and support.
4. Demonstrate management uses for data by sharing it with local actors and showing practical examples of how data can be used to improve service delivery.

Applying big data techniques, judiciously, in a data revolution, will be a key to faster human development and to better performance on the next generation of global goals.

* Specifically, the Eminent Persons Report says, "We also call for a data revolution for sustainable development, with a new international initiative to improve the quality of statistics and information available to citizens. We should actively take advantage of new technology, crowd-sourcing, and improved connectivity to empower people with information on the progress towards the targets" (p. 21).

† More information about PARIS21 (Partnership in Statistics for Development in the 21st Century) and the Center for Global Development is available from these organizations' Web pages, http://www.paris21.org/advocacy/informing-a-data-revolution and http://www.cgdev.org/.

RTI Press publication No. PB-0007-1506. Research Triangle Park, NC: RTI Press.
http://dx.doi.org/10.3768/rtipress.2015.pb.0007.1506

At the time the Eminent Persons Report[1] was released, the topic of big data as a corporate issue had already started to take off; a Google Trends search revealed a sharp increase in the usage of the term right around 2012. Big data became a frequent topic at meetings of non-data specialists in various institutions such as the World Bank and the United Nations, and in papers and books, especially regarding the implications for public policy.[4,5]

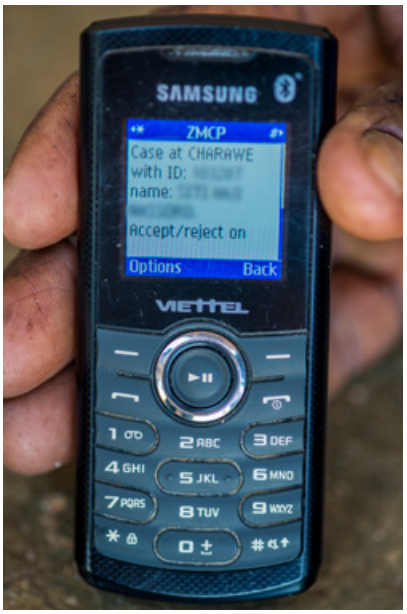
Although there are no rigorous, standard definitions of either "big data" or "data revolution for development," I provide two working definitions, and the rest of this brief both elaborates and distinguishes them further.

First, "big data" refers to the notion that data is, in itself, becoming a new economic resource that companies and public agencies can take advantage of. It gets its name from the fact that data flows have vastly increased in volume, velocity, and variety (the "three Vs"). Further, many people grasp that analyzing seemingly unrelated and unintentional data sets—such as scrutinizing Twitter feeds for public perceptions or to make predictions—can generate information with economic and social value. Several other features claimed to characterize big data are its emphasis on correlation rather than on cause and effect; the de-emphasis on careful, but also expensive, traditional sampling; and the making of a virtue out of inherent messiness (variety).

Second, "data revolution for development" has typically been understood as applying or even conflating the big data tenets described above to deal with development issues, mostly in the public sphere, in developing countries. This paper argues that while "borrowing" from big data and applying those borrowings to the development task may inspire better and more interesting uses of data, it may have some hidden dangers—among them the notion of minimizing the need for causal understandings.

As a related thought, many observers note that developing countries typically have not yet solved the more pedestrian "little data" issues: They may not know how many schools they have, or how many nurses are on payroll and how many are truly working. During a recent discussion on big data at the 2013 World Innovation Summit for Education, data specialist Emilio Porta made exactly this key point.[6]

This brief suggests how development agencies can help countries improve on the "little data" issue, partly by using modern but basic technology such as cell phones. At the same time, it points out how these countries can selectively and realistically apply some of the big data tenets to development issues.



The proliferation of cell phone technology in developing countries increases options for collecting data related to health, education, and other sectors of interest.

Big Data: Promise but Caution

Many of the popular big data tenets might be useful for the private sector in industrialized countries, or for multinational corporations in developing countries. And they could be inspirational for public policy improvement in developing countries. However, the tenets do not fully apply to key developing country, public management, and governance tasks, or at least to public-interest tasks such as improving education, health care, or food production. In particular, three of the ideas associated with big data require caution.

A first caution has to do with the notion that "correlation is enough; we do not really need to understand causality." An example of a correlation that might be useful in the private sector would be the 0.97 correlation between state-by-state beer sales and the numbers of teachers in United States. Suppose a country lacked population data, but had reasonably good data on teachers. This situation occurs in some countries—especially poor ones—because payroll systems are reasonably developed but censuses may lag or be purposefully perverted. In this hypothetical case, the number of teachers in that country would not be a bad proxy as an indicator with which to target, analyze, or predict beer sales. However, such correlations are not very useful for public policy on teaching or alcohol abuse. Thus, data revolution for development should be as concerned with causality as good empirical work on development should be.

A second cautionary note has to do with the idea that the data made possible by big data are universal (census-like), and therefore there is no need for careful thinking about sampling. Sampling techniques were developed because, traditionally, data were costly to gather, and there was a need to carefully

RTI Press publication No. PB-0007-1506. Research Triangle Park, NC: RTI Press.
http://dx.doi.org/10.3768/rtipress.2015.pb.0007.1506

sample subsets of populations in order to maximize the representativeness of the data. But in the literal sense, it would be impossible to take an absolute census of all members of a population at the same moment. In other words, all data sets are created by samples: A measurement taken one day is not the same as a measurement from another day or in different circumstances.

Moreover, self-selected reporting via crowd sourcing cannot produce data as relatively unbiased as good samples, especially if one needs to understand causality. And "unintentional data" may not even approach universality if, for example, smartphone and Internet penetration rates are low. Thus, in addition to consideration for causality, careful thinking about sampling is absolutely vital before data collection and analysis that might lead to high-stakes public policy.

The third caution is that in touting the advantages of variety, proponents of big data are, to some degree, making a virtue of the inherent messiness of unintentional data in unstructured formats. In situations where the need for causal analysis can be legitimately minimized, and where data sets arguably do apply to whole populations, touting messiness as a virtue may be defensible. But it is one thing to take advantage of variety; it is quite another to believe that vastly increasing the quantity of data and sources somehow removes the messiness.

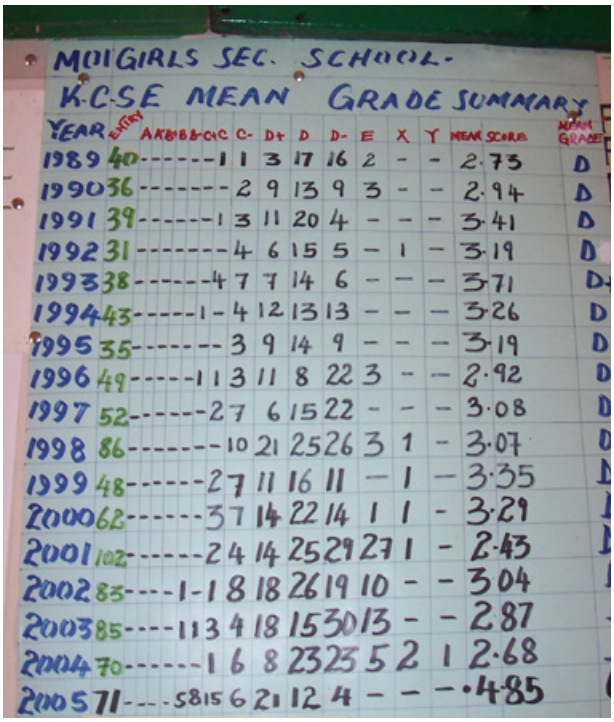
In a development and public policy context, key factors are clear thinking about data sources, limited amounts of data and data sources, good theory and causal models, and strong hooks to policy intent. Starting with this type of thinking may lead to collection of higher quality data than making do with inherently messy data.

A few more issues are pertinent. One problem in data revolution for development is not how to handle a torrent of data, but how to find any good data at all. Unlike within a company such as Walmart, data-based decision making in the public sphere of developing countries is hindered by the absence of the "little data" basic to administrative systems, such as knowing how many service points (schools, clinics) are in a country and their respective names and identification codes. And as mentioned above, in some countries even the trickle of existing data is not taken advantage of.

One distinction that is emerging between big data and data revolution for development has to do with the underlying motivations for the research. That is, the motive of big data among profit-driven production plants, service units, and firms is to generate a lot of data that could be transmuted into greater profits. In that sense, the demand for big data services is arising somewhat organically in the private sector, even if it is true that data scientists may have led the way in identifying the potential money hidden behind the use of data. In the case of data revolution for development, the need for data collection and analysis is driven more by a sense of professional responsibility, or doing public good, and is therefore felt by the relevant actors in a more diffuse and somewhat less pressing manner, especially in countries and public systems where accountability pressure is low anyway.

Issues in "Little Data" and Traditional Data Systems in Developing Countries

As suggested previously, current approaches and systems for managing any data at all—little or big—in the public sector of most developing countries leaves much to be desired. Data are often extremely slow to be gathered and are incomplete. As previously mentioned, central-level policy makers may not know the exact number of schools or clinics they have because no usable master list of service units exists, or the list is not accurate. For example, schools and clinics may have duplicate names or may not have a unique identifier that is used by all who gather data on that service unit. Many countries do not disaggregate data on outcomes at the citizen level (e.g., children's learning outcomes), and the data are not well curated (i.e., cleaned, sorted, and selected).



One school's posting of historical scores on the Kenya Certificate of Secondary Education.

RTI Press publication No. PB-0007-1506. Research Triangle Park, NC: RTI Press.
http://dx.doi.org/10.3768/rtipress.2015.pb.0007.1506

Typically, these "traditional" data systems were created mostly for reporting and, perhaps, top-down planning, whereas the current demands of development require more than that. Some of the resulting "flaws" that are particularly damaging are described below.

First, data sets are not often integrated. Although an education sector, for instance, may have data on student learning outcomes, these data are not integrated with data on school and community characteristics and resources used, among others. Thus, it is impossible (except in special studies) to determine the efficiency of each particular service unit (school, clinic, or local government's water utility) and whether it is high- or low-performing relative to expectations. It is also impossible to determine whether the service unit must be rewarded and emulated or whether it must be supported and pressured via accountability tools, including citizen-based accountability via voice or choice.

Second, data are not often linked to accountability mechanisms or accountability needs. Instead, the push to gather data and use them responds to particular stakeholders' professional and technical ideas on what types of information should be collected in order to do a better job. Or it may respond to overly simplistic views on planning, such as quantitative investment planning (how many clinics need to be built over the next ten years), which allows little space to answer to citizens, or to address the qualitative aspects of planning. The pressure to collect these data comes from donor agencies, from nongovernmental organizations, and sometimes from "transversal" ministries such as Finance or Planning, or from the more technical officials in line ministries. In some cases, the link to accountability is missing altogether, and hence the data systems are not really used or sustained.

Third, only rarely are data sourced directly from citizens—via mechanisms such as household surveys or direct requests for citizen feedback; further, as indicated in the first point above, systems that do collect such information may not integrate it with administrative data sets so that someone can act on it. At one extreme, the data system may contain no directly measured outcomes (i.e., only inputs and some outputs are measured), or no data on citizens' perceptions of service characteristics. Nor does information flow in the other direction, in the form of feedback on rights and service standards being provided to citizens.

In the absence of strong and clear service standards, high levels of citizen satisfaction on generic surveys will not necessarily reflect objective assessments of service quality and therefore will be unhelpful for policy making and planning. Ironically, citizens who are poor may be satisfied with low-quality services because they are unaware of service standards and lack the social power to impose implicit ones on the providers.

Toward a Proposed Approach: Tasks Required

Given the shortfalls of current approaches to data, and given the possibilities (and dangers) of generalizing from big data, I make four recommendations as a springboard to spur creative discussion and debate.

Improve "Little Data" and Add Citizen-Sourced Data

Improved and expanded reporting on standard administrative "little data," like that in legacy health or education management information systems (HMIS and EMIS), could contribute substantially to informed policy and program decision making.

Current donor support to countries emphasizes a lot of research and analytics, but neglects the relatively basic task of improving legacy systems. For instance, as previously mentioned, many countries do not have accurate lists of their total number of service units, or their databases do not match. Sectoral ministry registers on the numbers and locations of teachers, nurses, or extension agents often do not align with finance ministry records. "Ghost" teachers and nurses, and even whole "ghost" schools, abound. An important step to solve these disconnects would be to identify ways to triangulate these administrative data, especially by

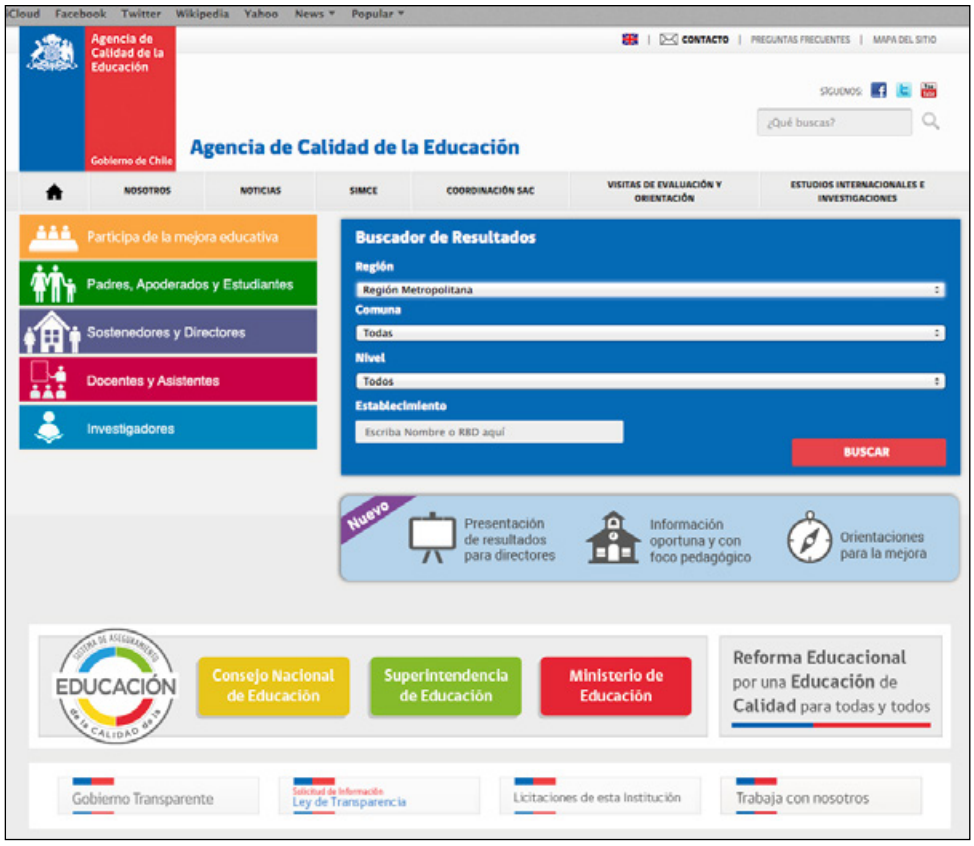
Mobile technology used for household-based malaria surveillance in Tanzania.

RTI Press publication No. PB-0007-1506. Research Triangle Park, NC: RTI Press.
http://dx.doi.org/10.3768/rtipress.2015.pb.0007.1506

(1) using even "dumb" cell phone technology as a way to double-check the "little data" and to improve system speed and completeness, and (2) adding citizen-sourced data. That said, data that continue to be obtained via traditional means also must be improved; the data could be supplemented by administering household surveys or by seeking information from, for example, heads of parent–teacher associations. Also important would be directly measuring citizen- or child-level outcomes (e.g., numeracy, freedom from waterborne diseases), not just outputs (e.g., attendance at school, access to water sources).

The use of data for citizen mobilization has been pioneered by organizations such as Twaweza (www.twaweza.org), which leads citizen-sourced efforts to measure the quality of outcomes of public services in East Africa. However, such efforts, while promising, have yet to demonstrate high impact; the challenge of involving citizens should not be underestimated. In addition, it may be necessary to combine these forms of citizen-oriented action with traditional bureaucratic action, as discussed below. More could be accomplished by adding technological components such as cell phones for data uptake or for feedback and mobilization of citizens.

Another improvement would be to move in the direction of universality (records for all receivers of services) and individuality (records all the way down to the individual level, not just from the clinic or the school). Obviously, privacy and rights issues must be respected, but this would not be an impossible task.



Chile's system for disseminating school results information.

A display on education finances at a Kenyan primary school, showing key inputs and outputs in a simple way.

RTI Press publication No. PB-0007-1506. Research Triangle Park, NC: RTI Press.
http://dx.doi.org/10.3768/rtipress.2015.pb.0007.1506

The key to improving "little data" and adding sound citizen-sourced data is to introduce as many types of accountability as possible. This might mean standard bureaucratic accountability; citizen or community accountability through voice, rights, and feedback; or market-based accountability through choice. Where data exist but are not used, the problem usually stems from a lack of decision-maker demand, which in turn results from weak accountability incentives and/or from problems in using existing data for accountability purposes. A data revolution for development should not encourage more supply-side data efforts similar to those carried out over the past decades that were not often sustained. A growing body of literature and experience reveals that engaging citizens in gathering data and using it for accountability can reinforce policy makers' and providers' performance incentives.[7,8]

Integrate, Curate, and Classify

Returning to one of the "flaws" indicated earlier, sectoral ministry officials and program managers rarely integrate existing data sets, much less complement existing data sets with new, citizen-sourced data. Views of the service units (e.g., schools, clinics, water utilities) are therefore partial.

For instance, some information about outcomes may be known, but nothing about costs—or vice versa. In some cases, managers may not perform the simple task of calculating cost per user and comparing across districts. As a result, one district might have double the input ratios of another district, but for no specific (technical or policy) reason. In research RTI carried out in Egypt, we found a case of one school district using 2,000 times as much photocopying budget per child as another district; we found this using the extremely simple process of calculating ratios.[9]

If systems were updated more regularly than yearly with data on single-unit or district efficiency, analysts could learn a lot from both the high- and low-performing units. Top performers could then be rewarded with formal merit pay or performance-based bonuses, or with more subtle schemes that relied on community esteem to support good performers. Low-performing service units could be given extra support or pressured to meet an acceptable standard. This type of attention to the data set is crucial to ensuring equity of resource and results distribution.

Add Analytical Value to Data Sets

If data are to be truly useful, something must happen to them to add value. Further, adding value should ideally go beyond vague attempts to use the data in management or planning. Even if the information is suited to, say, forecast the amount of inputs needed—as opposed to simply reporting numbers of persons served—that may be its weakest use. Even when actually carried out, such forms of adding value do not really address accountability issues or quality issues. Some ways to add value in a more meaningful manner are:

Push information from higher to lower levels. For inputs and outcomes, generate data that enable ready comparisons across health facilities and districts, school districts, and/or municipalities. This requires sending information back down to the local level, as opposed to using traditional systems in which data go from the bottom to the top of the administrative hierarchy and (presumably) feed only sectoral policy and planning decisions. Some examples include providing comparative data on service standards and ratios to citizens, boards, users' groups, and districts. Comparisons of results and costs and of input ratios (e.g., teacher–pupil ratios) across jurisdictions are an excellent way to spur discussions among citizens about the data and to create a demand for the data, because they link to basic concepts of fairness and efficiency. The link to citizen-sourced satisfaction or service perception data can be leveraged.

Prepare predictive analytics or simple forms of a "positive deviant" analysis. The results can then be used to reward and emulate high-performing service units or to apply pressure and support low-performing units. Another method is to promote stories and case studies about low-performing service units and how they improved their performance.

Link to highly specific forms of data usage, so as to demonstrate the power of data. Thus, for instance, instructional coaches' records from teacher observations and feedback could be merged with average reading results for students in those teachers' classrooms. Data on observation of work processes in the health or water sectors could be linked to improvements in illness outcomes and municipal water supply parameters. Specificity is key because data usage "shines" most when it is tied to highly specific outcomes or inputs, as opposed to general outcomes such as citizen satisfaction or broad sectoral coverage targets. This is because the linkages among measurement, response, and re-measurement become more evident. For example, did malaria rates decrease in particular locations as compared to others, and do the citizens know about it? Are the teachers absent or present? Are children's stunting rates in particular locations decreasing?

Connect the data findings to a library of solutions. The data derived from data revolution sorts of interventions—comparative data, citizen-fed data, data on positive and negative deviants—can be linked to prototypical solutions. For example, service units might be able to learn from standardized approaches to improved facility management, based on empirical evidence from those that are managed well. Another stock of knowledge

RTI Press publication No. PB-0007-1506. Research Triangle Park, NC: RTI Press.
http://dx.doi.org/10.3768/rtipress.2015.pb.0007.1506

may be randomized controlled trials and other forms of rigorously evaluated interventions. This information can be merged with what is learned via the case-specific study of positive deviants previously discussed. These libraries of solutions can help improve the stock of support offered to low-performing units, such as better training for nurses on monitoring children's growth, or for waterworks employees on controlling leaks. Significantly, merely pressuring underperforming units with comparative data is unlikely to lead to improvements unless the comparative data are paired with support on *how* to improve—a notion which leads to the last proposed step.

Demonstrate Use in Management and in Accountability and Governance

A key component of the data revolution for development is to go beyond data production, curation, integration, and adding analytical value and to use data to drive performance and management improvements in schools, clinics, and other production units. Closing this loop and connecting directly to decision-making constitutes a separate set of activities that need to be planned.

Electronic tablets are used for collecting and storing data locally, to be synchronized with a remote server when a network connection is available.

A number of countries are experimenting with the use of information technology to facilitate the link to service improvement and better performance. For example, providing programmed electronic tablets to site inspectors and coaches (of nurses, teachers, water-user group workers) could enable them to input data and to access a menu of coaching advice. In some cases, the tablets could provide video clips on recommended actions. Another application could involve getting these same coaches to actively demonstrate the use of the data in public or private meetings where schools, clinics, water facilities, and other service provider units are subject to accountability pressures. To solidify the link to managing for performance, information systems need to track, document, and evaluate not just whether, but also how, the use of data leads to improvement.

Our research suggests that highly specific actions are needed to have a sharp demonstration effect—that is, not just improving education quality, but improving children's learning of reading in grades 1–3, perhaps.[10] An even simpler but highly specific intervention, facilitated by the use of tablets and community vigilance, would be around guaranteeing actual teaching time. However, the effects of these actions are typically limited to the targeted behaviors or skills. Highly specific and data-based support actions that can get nurses to improve on, say, growth monitoring or oral rehydration therapy will have the intended effect, and therefore can demonstrate the impact of using data to drive performance improvement, but will not *automatically* spill over into generalized improvement in child welfare services. The demonstrated feedback loop from data to performance improvement can be generalized to many indicators, so one does not have to have demonstration pilots for every single behavior. However, it has to be generalized in a conscious way.

Conclusions

Donor agencies and countries should create projects to test and then put to use the principles outlined in this brief. This will not be inexpensive. But, as the Post-2015 Copenhagen Consensus process has noted, not all indicators being proposed for the Sustainable Development Goals should have equal priority.[11] Essential will be picking relatively high-value, easy, impactful indicators for the most important goals—those around early childhood well-being, for instance—and then using these to test and apply the principles in this brief.

RTI Press publication No. PB-0007-1506. Research Triangle Park, NC: RTI Press.
http://dx.doi.org/10.3768/rtipress.2015.pb.0007.1506

References

1 United Nations. A new global partnership: eradicate poverty and transform economies through sustainable development. The report of the high-level panel of eminent persons on the post-2015 development agenda. New York: United Nations; 2013 [cited 2015 May 06]. Available from: http://www.un.org/sg/management/pdf/HLP_P2015_Report.pdf

2 World Bank. Big data in action for development. Washington, DC: World Bank; 2014 [cited 2015 May 06]. Available from: http://live.worldbank.org/sites/default/files/Big%20Data%20for%20Development%20Report_final%20version.pdf

3 Center for Global Development and African Population and Health Research Centre. Delivering on the data revolution for Sub-Saharan Africa: final report of the Data for African Development Working Group. Washington, DC: Center for Global Development; 2014 [cited 2015 May 06]. Available from: http://www.cgdev.org/sites/default/files/CGD14-01%20complete%20for%20web%200710.pdf

4 Cukier KN, Mayer-Schoenberger V. The rise of big data: how it's changing the way we think about the world. Foreign Aff; 2013; May–June. Available from: https://www.foreignaffairs.com/articles/2013-04-03/rise-big-data

5 Zikopoulos PC, Eaton C, deRoos D, Deutsch T, Lapis G. Understanding big data: analytics for enterprise class Hadoop and streaming data; 2012. New York: McGraw-Hill.

6 World Innovation Summit for Education (WISE) 2013. Debate: what data for what purpose? [cited 2015 May 07]. Video: mins 21:37-23:26. Doha, Qatar, Oct 29–31, 2013, summit theme "Reinventing Education for Life." Available from: https://www.youtube.com/watch?v=2bX0McIVQoE

7 World Wide Web Foundation [Internet]. Research project: exploring the emerging impacts of open data in developing countries [cited 2015 May 07]. Funded by the Canadian International Development Research Centre, grant no. 107075. Washington, DC: World Wide Web Foundation. Available from: http://www.opendataresearch.org/project/2013/oddc

8 Schwegmann C. Open data in developing countries. European Public Sector Information Platform (ePSI*platform*) topic report no. 2013/02 [cited 2015 May 07]. Available from: http://www.epsiplatform.eu/sites/default/files/127790068-Topic-Report-Open-Data-in-Developing-Countries.pdf

9 Healey FH, Crouch LA, Hanna R. Formula-based decentralization of non-personnel, non-capital ("Bab 2") spending in Egyptian education: problems and possible solutions. Technical report prepared for USAID under the EQUIP2 Education Reform Project in Egypt. Research Triangle Park (NC): RTI International; 2010.

10 Piper B, Zuilkowski, SS, Mugenda A. Improving reading outcomes in Kenya: first-year effects of the PRIMR Initiative. Int J Educ Dev. 2014; 37:11–21. http://dx.doi.org/10.1016/j.ijedudev.2014.02.006

11 Copenhagen Consensus Center. Post-2015 consensus: what are the smartest targets for the post-2015 development agenda? [Internet] [cited 2015 May 07]. Available from: http://www.copenhagenconsensus.com/post-2015-consensus

About the Author

Luis Crouch, (PhD, Agricultural Economics, University of California, Berkeley) is Chief Technical Officer within the International Development Group at RTI International. He has provided policy advice on data and finance to many governments around the world, has a long record of using data for policy advice, and has written extensively on the use of data for development planning and service improvement.

RTI Press Research Briefs and Policy Briefs are scholarly essays on policy, methods, or other topics relevant to RTI areas of research or technical focus.

RTI International, 3040 East Cornwallis Road, PO Box 12194
Research Triangle Park, NC 27709-2194 USA

+1.919.541.6000 rtipress@rti.org www.rti.org

www.rti.org/rtipress