WPS5532

Policy Research Working Paper 5432

Mashup Indices of Development

Martin Ravallion

The World Bank
Development Research Group
Director's office
September 2010

Policy Research Working Paper 5432

Abstract

Countries are increasingly being ranked by some new "mashup index of development," defined as a composite index for which existing theory and practice provides little or no guidance to its design. Thus the index has an unusually large number of moving parts, which the producer is essentially free to set. The parsimony of these indices is often appealing—collapsing multiple dimensions into just one, yielding unambiguous country rankings, and possibly reducing concerns about measurement errors in the component series. But the meaning, interpretation and robustness of these indices are often unclear. If they are to be properly understood and used, more attention needs to be given to their conceptual foundations, the tradeoffs they embody, the contextual factors relevant to country performance, and the sensitivity of the implied rankings to changing the data and weights. In short, clearer warning signs are needed for users. But even then, nagging doubts remain about the value-added of mashup indices, and their policy relevance, relative to the "dashboard" alternative of monitoring the components separately. Future progress in devising useful new composite indices of development will require that theory catches up with measurement practice.

This paper—a product of the Director's office, Development Research Group—is part of a larger effort in the department to assess development indicators. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The author may be contacted at mravallion@worldbank.org.

Mashup Indices of Development

Martin Ravallion[1]

Development Research Group, World Bank
1818 H Street NW, Washington DC, 20433, USA

[1] For helpful comments the author is grateful to Sabina Alkire, Kathleen Beegle, Rui Manuel Coutinho, Asli Demirguc-Kunt, Quy-Toan Do, Francisco Ferreira, Garance Genicot, Carolin Geginat, Stephan Klasen, Steve Knack, Aart Kraay, Will Martin, Branko Milanovic, Kalle Moene, Dominique van de Walle, Roy Van der Weide and Hassan Zaman. These are the views of the author, and need not reflect those of the World Bank or any affiliated organization.

Various indicators are used to track development, both across countries and over time. The World Bank's annual *World Development Indicators* presents literally hundreds of development indicators (World Bank, 2009). The UN's *Millennium Development Goals* are defined in terms of multiple indicators. Even in assessing specific development goals, such as poverty reduction, mainstream development thinking and practice is premised on a multidimensional view, calling for a range of separate indicators.

Faced with so many indicators—a "large and eclectic dashboard" (Stiglitz et al., 2009, p.62)—there is an understandable desire to reduce the dimensionality, to form a single composite index. As Samuelson (1983, p. 144) put it (in the context of aggregating commodities): "There is nothing intrinsically reprehensible in working with such aggregate concepts." However, as Samuelson goes on to note in the same passage: "… it is important to realize the limitations of these aggregates and to analyze the nature of their construction."

Two broad types of composite indices of development can be identified. In the first, the choices of the component series and the aggregation function are informed and constrained by a body of theory and practice from the literature. GDP, for example, is a composite of the market values of all the goods and services produced by an economy in some period. Similarly, aggregate consumption is a composite of expenditures on commodities. A standard poverty or inequality measure uses household consumption or income, which are aggregates across many components. In these cases, the composite index is additive and linear in the underlying quantities, with prices (including factor prices) as their weights. A body of economics helps us construct and interpret such indices. With a complete set of undistorted competitive markets, market prices are defensible weights on quantities in measuring national income, though even then we will need to discount this composite index for the extent of income inequality to derive an acceptable money metric of social welfare (under standard assumptions). And market prices will need to be replaced by appropriate shadow prices to reflect any market imperfections such as rationing. There is a continuing debate and reassessment related to these and other aspects of measurement, through which practice gets refined. Decisions about measurement are guided by an evolving body of theory and practice.

This is not the case for the second type of composite index. Here the analyst identifies a set of indicators that are assumed to reflect various dimensions of some unobserved (theoretical)

concept. An aggregate index is then constructed at the country level, usually after re-scaling or ranking the component series.[2] Neither the menu of the primary series nor the aggregation function is pre-determined from theory and practice, but are "moving parts" of the index—key decision variables that the analyst is free to choose, largely unconstrained by economic or other theories intended to inform measurement practice.

Borrowing from web jargon, the data going into this second type of index can be called a "mashup." In web applications one need not aggregate the data into a composite index; often users look instead for patterns in the data. When a composite index is formed from the mashup, I will call it a "mashup index." This is defined as a composite index for which the producer is only constrained by the availability of data in choosing what variables to include and their weights.

To illustrate the distinction, consider two stylized examples of composite indices, both formed from the data on household assets and consumer durables found in the *Demographic and Health Surveys* (DHS). For index A the variables and their weights are set by the analyst, who has some concept of "economic welfare" in mind, and thinks this is related to certain variables in the DHS, which are aggregated based on the analyst's judgments. For index B, the variables and weights are instead based on a regression model calibrated to another survey data set for which a comprehensive measure of consumption (though still containing measurement errors) could be derived. The model is calibrated to common variables in the expenditure survey and the DHS, and the regression model is used to predict wealth in the DHS. A is a mashup index, B is not.

The country rankings implied by mashup indices often attract media attention. People are naturally keen to see where their country stands. However, the details of how the composite index was formed—the variables and weights—rarely get the same scrutiny. Typically the (often web-based) publications do not comply with prevailing scholarly standards for documenting and defending a new measure. No doubt many users think the index has some scientific status.

Just as it is recognized that there can be gains from bringing together data and functionality from different sources in creating a web-application hybrid, there can be gains in

[2] A common re-scaling method is to normalize the indicator x to be in the (0,1) interval by taking the transformation $(x-\min(x))/(\max(x)-\min(x))$ where $\min(x)$ is the lowest value of x in the data and $\max(x)$ is the highest value, and then add up the re-scaled indicators. The most common ranking method is to rank countries by each indicator x and then derive an overall ranking according to the (weighted) aggregate of the rakings across components (a version of the voting method called the Borda rule).

forming a mashup index. These gains often stem from the inadequacies of prevailing composite indices of the first type as characterizations of important development goals—combined with the desire for a single (scalar) index. As data sources become more open and technology develops, creative new mashups can be expected. It is a good time then to take stock of the concerns with existing indices, in the hope of doing better in the future.

This paper offers a critical assessment of the strengths and weaknesses of existing mashup indices of development. One theme of the paper is the importance of assessing the (rarely explicit) tradeoffs embodied in these indices—for those tradeoffs have great bearing on both their internal validity and their policy relevance. Another theme is the importance of transparency about the robustness of country rankings. Clearer warnings are needed for users, and technology needs to be better exploited to provide those warnings. As it is, prevailing industry standards in designing and documenting mashup indices leave too many things opaque to users, creating hidden costs and downside risks, including the diversion of data and measurement efforts, and risks of distorting development policy making.

After describing some examples, the paper discusses the generic questions raised by mashup indices.

Examples of mashup indices of development

A prominent set of examples of mashup indices is found in past efforts to combine multiple social indicators. An early contribution was the *Physical Quality of Life Index* (PQLI) (Morris, 1979), which is a weighted average of literacy, infant mortality and life expectancy. Along similar lines, a now famous example is the *Human Development Index* (HDI) that is published each year in the UNDP's *Human Development Report* (HDR), which started in 1990. The HDI adds up attainments in three dimensions—life expectancy, schooling (literacy and enrollment rates) and log GDP per capita at purchasing power parity—after re-scaling each of them.[3] There have been a number of spinoffs from the HDI, including the "*Gender Empowerment Measure*," which is a composite of various measures of gender inequalities in

[3] See Anand and Sen (2000) for a useful overview of the construction of the HDI and how this has changed over time. The 2010 HDR introduced some further changes to the variables and aggregation function. I will comment on these changes later.

political participation , economic participation and decision making, and power over economic resources.

In a similar spirit to the HDI, the *Multidimensional Poverty Index* (MPI) was developed by Alkire and Santos (2010a), in work done for the 2010 HDR. The authors choose 10 components for the MPI; two for health (malnutrition, and child mortality), two for education (years of schooling and school enrolment), and six aim to capture "living standards" (including both access to services and proxies for household wealth). Poverty is measured separately in each of these 10 dimensions, each with its own weight. In keeping with the HDI, the three main headings—health, education, and living standards—are weighted equally (one-third each) to form the composite index. A household is identified as being poor if it is deprived across at least 30% of the weighted indicators. While the HDI uses aggregate country-level data, the MPI uses household-level data, which is then aggregated to the country level. Alkire and Santos construct their MPI for more than 100 countries.[4]

Mashups have been devised for other dimensions of development. The "*Economic Freedom of the World Index*" is a composite of indices of the size of government, property rights, monetary measures (including the inflation rate and freedom to hold foreign currency accounts), trade openness and regulation of finance, labor and business (Gwartney and Lawson, 2009). The "*Worldwide Governance Indicators*" (WGI) (Kaufmann, Kraay and Mastruzzi, 2009), is a set of mashup indices, one for each of six assumed dimensions of governance: voice and accountability, political stability and lack of violence or terrorism, governmental effectiveness, regulatory quality, rule of law, and corruption. The WGI covers some 200 countries and is now available for multiple years.

Probably the most well-known mashup index produced by the World Bank Group is the "*Ease of Doing Business Index*"—hereafter the Doing Business Index (DBI).[5] This is a simple average of country rankings for ten indices aiming to measure how easy it is to open and close a business, get construction permits, hire workers, register property, get credit, pay taxes, trade across borders and enforce contracts. Unlike most of the mashup indices, DBI collects its own

[4] See Ravallion (2010a) for further discussion of multidimensional indices of poverty, including the MPI.

[5] This developed from an original data compilation documented in Djankov et al. (2002).

data, using 8,000 local (country-level) informants. The composite index is currently produced for 183 countries. The country rankings are newsworthy, with over 7,000 accumulated citations in Google News.

The World Bank's "*Country Policy and Institutional Assessments*" (CPIA) attempt to assess the quality of a country's policy and institutional environment. This is not a mashup index, but it is used to produce what is arguably the most important of any of the mashup indices of development. The CPIAL has 16 components in four clusters: economic management (macro management, fiscal and debt policies), structural policies (trade, finance, business and regulatory environment), policies for social inclusion and equity (gender equality, human resources, social protection, environmental sustainability) and governance (property rights, budgetary management, revenue mobilization, public administration, transparency and accountability in public sector). These are all based on "expert assessments" made by the Bank's country teams, who prepare their proposed ratings, with written justifications, which are then reviewed.

Two mashup indices are produced from the CPIA. One of them is simply an equally weighted sum of the four cluster-specific indices, with equal weights on their sub-components. This appears to be only used for presentational purposes. The second index puts a weight of 0.68 on the governance cluster of the CPIA and 0.24 to the mean of the other three components (and the remaining weight goes to the Bank's assessment of the country's "portfolio performance"). This "governance-heavy" mashup index based on the CPIA is used in allocate the World Bank's concessional lending, called "International Development Association" (IDA), across IDA eligible countries. The African Development Bank has undertaken a similar CPIA exercise to guide its aid allocation decisions.

The *Environmental Performance Index* (EPI), produced by teams at Columbia and Yale Universities, is probably the most well known mashup index of environmental data. This ranks 163 countries by a composite of 25 component series grouped under 10 headings: climate change, agriculture, fisheries, forestry, biodiversity and habitat, water, air pollution (each of the latter two having two components, one for effects on the ecosystem and one for health effects on humans) and the environmental burden of disease.

Probably the most ambitious example yet of a mashup using development data was released by Newsweek magazine in August 2010. This tries to identify the "*World's Best Countries*," using a composite of many indicators (many of them already mashup indices) assigned to five groupings: education, health, quality of life, economic competitiveness and political environment. The education component uses test scores. The health component uses life expectancy at birth. "Quality of life" reflects income inequality, a measure of gender inequality, the World Bank's poverty rate for $2 a day, consumption per capita, homicide rates, the EPI, and the unemployment rate. "Economic dynamism" is measured by the growth rate of GDP per capita, non-primary share of GDP, the World Economic Forum's Innovation Index, the DBI and stock market capitalization as a share of GDP. The "political environment" is measured by the Freedom House ratings, and measures of political participation and political stability.

The rest of this paper critically reviews the main claims made about the benefits of these and other mashup indices of development. Rather little seems to be known about their costs. The teams working on these indices appear to range from just a few people to 30 or more. The web site for Doing Business lists 33 staff on the team who produced the 2010 edition, on top of the 8,000 "local experts."[6] However, it should be recalled that this team is collecting the primary data, so this does not imply a high cost of the mashup index *per se*. The labor inputs to producing prevailing mashup indices are probably small.

Questions to ask about any mashup index

One can readily sympathize with the motivation for a mashup index. No single data series captures the thing one is interested in, so by adding up multiple indices one may hope to get closer to that truth; in principle there can exist an aggregate index that is more informative than any of its components. It is another matter whether this sympathy survives a closer inspection of what is done in practice. What goes into the mashup and how useful is what comes out?

Four main issues can be identified: the need for conceptual clarity on what is being measured, the need for transparency about the tradeoffs embedded in the index, the need for robustness tests and the need for a critical perspective on policy relevance. These are not solely

[6] The DBI project does not apparently pay these local experts, though, of course, their time has value, and so it should be included in assessing the full cost of the DBI.

issues for mashup indices; practices for other composite indices are often less than ideal in these respects. However, by their very nature—as composite indices for which virtually everything is up for grabs—these concerns loom especially large for mashup indices.

What is being measured and why? The fact that the target concept is unobserved does not mean we cannot define it and postulate what properties we would like its measure to have. Understanding the purpose of the index can also inform choices about its calibration.

In practice we are often left wondering what the concept is that the index is trying to measure and why. For example, what exactly does it mean to be the "best country" in Newsweek's rankings (which turns out to be Finland). (I guess I should be pleased to see my country, Australia, coming in at number 4, but I have little idea what that means.) The rationale for the choices made in the Newsweek index is far from clear, not least because one is unsure what exactly the index is trying to measure.[7]

Some mashup indices have been motivated by claimed inadequacies in more standard development indices. The construction of a number of the mashup indices of development has been motivated by the argument that GDP is not a sufficient statistic for human welfare—that it does not reflect well the concerns about income distribution, sustainability and human development that matter to welfare. To my eyes this is a straw man, and it has been so for a long time. Soon after the HDI first appeared, motivated by these inadequacies of GDP, Srinivasan (1994, p.238) wrote: "In fact, income was never … the sole measure of development, not only in the minds of economists but, more importantly, among policy makers." In poverty measurement, a similar straw man is the view that mainstream development thinking has been concerned solely with "income-poverty," ignoring other dimensions of welfare. For example, in Alkire and Santos (2010b), the authors of the MPI counterpoint their measure with the World Bank's "$1 a day" poverty measures, which use household consumption of commodities per

[7] Why, for example, does "economic dynamism" matter independently of the standard of living in the Newsweek index? The way we normally think about this, it is not economic growth *per se* that helps deliver human welfare but the realized level of living. But maybe there is some other concept of what it means to be the "best country" that motivated this choice such as the possibility of being the best country at some time in the future. There are also some puzzles in the choices made for filling in missing data; for example, for some unexplained reason a "Global Peace Index" was used for the Gini index of inequality when the latter was missing. Greater conceptual clarity might also help guide such choices.

person as the metric for defining poverty.[8] Yet, while it is true that the World Bank puts considerable emphasis on the need to reduce consumption or income poverty, it is certainly not true that human development is ignored; indeed, this topic has a prominent place in the Bank's work program, side-by-side with its focus on income poverty.[9] A similar comment can be made regarding environmental sustainability, which has a prominent place in the Bank's work.

The fact that a welfare indicator is in monetary units cannot be objectionable *per se*. One could in principle construct a money-metric of almost any agreed (multidimensional but well-defined) welfare concept. A strand of the economics literature on welfare measurement has taken this route, by deriving money metrics of welfare from an explicit formulation of the individual and social welfare functions.[10] Conventionally, those functions have been seen to depend on command over commodities (allowing for inequality aversion), but the approach can be extended to important "non-income" dimensions of welfare. For example, Jones and Klenow (2010) introduce life expectancy into a money metric of social welfare (embodying inequality aversion) based on expected utilities, where life expectancy determines the probability of realizing positive welfare (with utility scaled to be zero at death). Arguably the important issue is not the use of a monetary metric, but whether one has used the right components and prices in evaluating that metric.

Some mashup indices have alluded to theoretical roots, to help give credibility. However, there is invariably a large gap between the theoretical ideal and what is implemented. For example, the HDI claims support from Sen's writings arguing that human capabilities are the relevant concept for defining welfare or well-being (see, for example, Sen, 1985). Yet it is quite unclear how one goes from Sen's relatively abstract formulations in terms of functionings and capabilities to the specific mashup index that is the HDI. Why, for example, does the HDI include GDP, which Sen explicitly questions as a relevant space for measuring welfare?[11] Sen

[8] The latest update described in Chen and Ravallion (2010).

[9] The Bank devotes a great deal of attention to the measurement of health and education attainments and the quality of public services as part of its Human Development Vice-Presidency and its Human Development and Public Services division within the research department.

[10] For example, under certain conditions a money metric of aggregate social welfare can be derived by deflating national income by appropriate social cost of living indices; for a good overview of this literature see Slesnick (1998).

[11] Presumably in response to this question, more recent HDRs have provided a "non-income HDI" that exclude GDP per capita. However, the bulk of attention goes to the ordinary HDI. Anand and Sen (2000)

has also questioned whether life expectancy is a good indicator of the quality of life; Sen (1985, p.30) notes that "The quality of life has typically been judged by such factors as longevity, which is perhaps best seen as reflecting the quantity (rather than quality) of life." Possibly it is the combination of GDP and life expectancy that somehow captures "capabilities," but then where in Sen's writings do we find guidance on the valuation of life, as required by any (positively weighted) aggregation function defined on income and life expectancy? (I return to the issue of tradeoffs below.) It is clearly a large step indeed from Sen's (often powerful) theoretical insights to the idea of "human development" found in the HDRs, and an even bigger step to the specific measure that is the HDI.

A similar comment applies to the MPI. In defending their data and methodological choices, the authors of the MPI contrast their index to poverty measures based on consumption or income, arguing that "the MPI captures direct failures in functionings that Amartya Sen argues should form the focal space for describing and reducing poverty" (Alkire and Santos, 2010a, p.1). However, the various components of the MPI include measures of deprivation in the attainments space as well as functionings. As with the HDI, it is unclear how much this really owes to Sen. And if one looks at how poverty lines are in fact constructed for most conventional poverty measures found in practice, they too can claim no less credible antecedents in Sen's approach. By this interpretation, the poverty line is the monetary cost of attaining certain basic functionings, as outlined in Ravallion (2008). In practice, the main functioning is adequate nutritional intakes for good health and normal activities, though an allowance for basic non-food needs is almost always included. More generally one can define a poverty line as a money metric of welfare. By normalizing consumption or income by such a poverty line,[12] the resulting poverty measure comes to reflect something closer to the broader concept of welfare than the authors of the MPI appear to have in mind. The key point here is that doing analysis in the income space does not preclude welfare being defined in other spaces, as has long been recognized in economics.

---

discuss specifics of how GDP per capita enters the HDI. (The income variable switched to Gross National Income in the 2010 HDR.)

[12] Blackorby and Donaldson (1987) call these "welfare ratios" and show that aggregating empirical money-metric welfare ("equivalent income") functions into empirical social welfare functions can be problematic unless the money metric of utility can be written as a welfare ratio.

In truth, the concept of "human development" in the HDI has never been crystal clear and nor is it clear how one defines the broader concept of "poverty" that indices such as the MPI are trying to capture, and how this relates to "human development." Development policy dialogues routinely distinguish "poverty" from "human development," where the poverty concept relates to command over commodities. While "poverty" is typically distinguished from "human development," it can be argued that mainstream development thinking and practice is already premised on a multidimensional view of poverty (Ravallion, 2010a). The real issues are elsewhere, in the case for and against forming a mashup index.

The frequent lack of conceptual clarity about what exactly one is trying to measure makes it hard to judge the practical choices made about what pre-existing indicators get used in the composite. One can debate the precise indicators chosen, as would probably always be the case. Double counting is common,[13] though unavoidable to some degree. But greater guidance for users on the properties of the ideal measure with perfect data would help assess the choices made with imperfect data. For example, while we can agree that "income" (as conventionally measured) is an incomplete metric, we would presumably want any measure of "poverty" to reflect well the changes in peoples' real incomes (their command over commodities)—changes that might emanate from shocks. The MPI's six "living standard" indicators are likely to be correlated with consumption or income, but they are unlikely to be very responsive to economic fluctuations. The MPI would probably not capture well the impacts on poor people of the Global Financial Crisis, or rapid upswings in macro-economic performance.

What tradeoffs are embedded in the index? We need to know the tradeoffs—defined here as the marginal rates of substitution (MRS)[14]—built into a composite index if it is to be properly assessed and used. If a policy or economic change entails that one of the positively-valued dimensions of welfare increases at the expense of another such dimension, then it is the MRS that determines whether overall welfare has risen or fallen. However, whether or not one

[13] For example, private and public spending on health and education is a component of GDP, while measures of health and education attainments also enter separately in the HDI. In the case of the Newsweek index, mean consumption enters both directly (on its own) and indirectly via other variables, notably the poverty rate, which is also a function of inequality, which also enters on its own.

[14] Consider any (differentiable) function f of x_1, x_2. The MRS of $f(x_1, x_2)$ is simply the ratio of the first derivative ("weight") with respect to x_1 divided by the first derivative with respect to x_2. This gives how much extra x_2 is needed to compensate for one unit less of x_1, where "compensate" is defined as keeping the value of $f(x_1, x_2)$ constant. (More general definitions are possible without assuming differentiability.)

thinks that a mashup index has some status as a policy objective, knowing its weights and (hence) tradeoffs is key to understanding the properties of the index.

At one level, the weights in most mashup indices are explicit.[15] Common practice is to identify a set of component variables, group these in some way, and attach equal weight to these groups for all countries.[16] It is hard to believe that weights could be the same for all countries, and (indeed) all people within a country. Unlike market prices, which will come into at least rough parity within specific economies (and between countries for traded goods), the values attached to non-market goods will clearly vary with the setting, including country or individual attributes. For example, the weight attached to access to a school will depend on whether the household has children. The weights attached to the various dimensions of good policies and institutions identified in the CPIA surely cannot be the same in all countries, as critics have noted.[17]

There are typically two levels at which weights can be defined in mashup indices. First there are the (typically equal) weights on the components indices, such as "education," "health" and "income" in the HDI. However, the component indices are invariably functions of one or more primary variables (such as literacy and school enrollment in the education component of the HDI). While the weights attached to the component indices are typically explicit, this is almost never the case for the weights attached to the underlying dimensions. The explicit weights are defined in an intermediate, derived, space. Indeed, little or no attention is given to the implied tradeoffs in the space of the primary dimensions being aggregated, and whether they are defensible. It does not even appear to be the case that the aggregation functions in most of the current mashup indices of development have been chosen with regard to the implied tradeoffs on those dimensions.

---

[15] Stiglitz et al. (2009) note approvingly that popular composite indices use explicit weights. Nonetheless, the weights can remain opaque in the most relevant space for user assessment. The tradeoffs in those dimensions can also be crucial to the "normative implications," which are often unclear for prevailing composite indices, as Stiglitz et al. (2009) also point out.

[16] For example, the health, education and income components of the HDI get equal weight, similarly to the MPI, and the EPI gives equal weight to environmental impacts on the ecosystem and human health.

[17] See the discussion of the "Performance Based System" (which includes the CPIA) in African Development Bank (2007, Chapter 4).

For those indices (such as DBI) that are created by taking an average of the rankings of countries by the components, it is quite unclear what the weights are on those components; the mean rank is typically equally weighted, but the weights on any primary variable—the first derivative of the composite index with respect to that variable—are unknown, and difficult to determine. There can be no presumption that the MRS would have seemingly desirable properties; using this method of aggregation it is possible that a component that has a low value in some country will not be valued highly relative to another component with a high value. In other words, the MRS need not decline as one increases one component at the expense of the other.[18] These aggregation methods are thus capable of building in perverse valuations.

In some cases one can figure out the implicit tradeoffs, even though they are not explicit in the documentation of the mashup index. The tradeoffs embodied in the HDI have been particularly contentious in the literature.[19] By adding up average income per capita with life expectancy (after re-scaling and transforming each component) the HDI implicitly puts a monetary value on an extra year of life, and that value is deemed to be much lower for people in poor countries than rich ones.[20]

Figure 1 gives the extra income that would be needed to compensate for one year less life expectancy implicit in the 2010 version of the HDI. The value of life varies from very low levels in poor countries—the lowest value of $0.50 per year is for Zimbabwe—to almost $9,000 per year in the richest countries. Granted Zimbabwe is an outlier, even amongst low-income countries; the next lowest is Liberia, with a value of $5.51 per year attached to an extra year of life. However, the same point remains: the HDI implicitly puts a much lower value to extra life in poorer countries than rich ones.[21]

---

[18] This is easy to see if one assumes that the number of countries is large and the component variables have continuous distributions, with smooth unimodal densities (such as normal densities). The MRS between two components of a composite index based on average ranks will then be the relative probability densities and it is plain that the curvature of the implied contours is theoretically ambiguous.

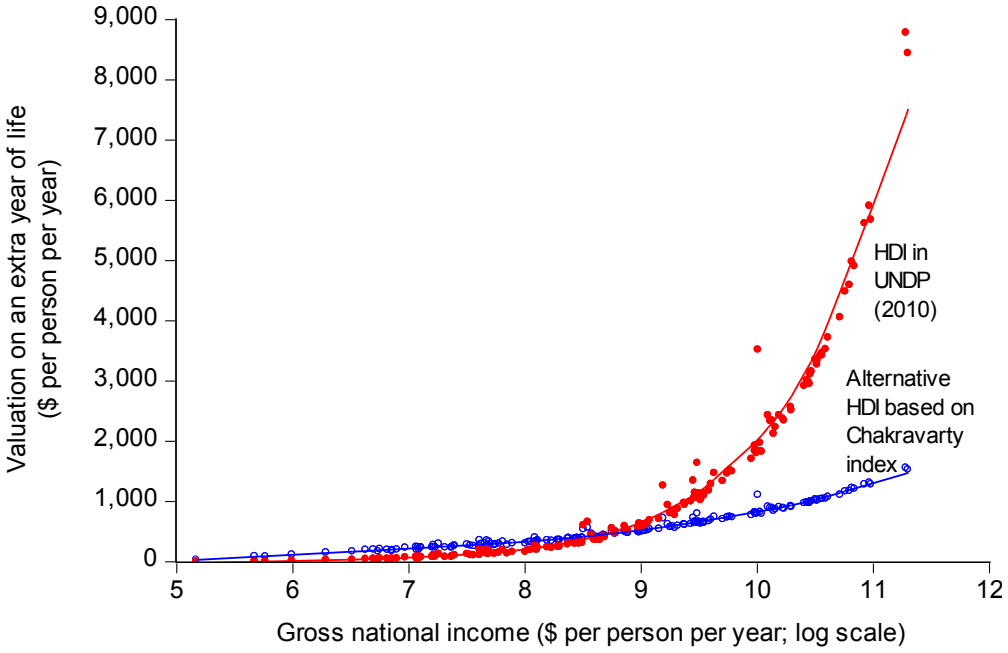
[19] Contributions on this issue include Kelley (1991), Ravallion (1997) and Segura and Moya (2009).

[20] This was first pointed out by Ravallion (1997). The 2010 HDI's implicit MRS (the extra income needed to compensate for a year less life expectancy) is $Y(ln\,Y - lnY^{min})/(LE - LE^{min})$ where Y is Gross National Income per capita, *LE* is life expectancy at birth and the "min" values are set by the 2010 HDR at $163 per year and 20 years for Y and *LE* respectively.

[21] For further discussion of the implicit tradeoffs built into the HDI and how they have changed see Ravallion (2010b).

A rich person will clearly be able to afford to spend more to live longer than a poor person, and will typically do so. But should we build such inequalities into our assessment of a country's progress in "human development"? Does the HDR really want to suggest that, in the interests of promoting "human development," Zimbabwe should not be willing to implement a policy change that increases life expectancy by (say) one year if it lowers national income per capita by more than $0.51—barely 0.3% of the country's income? Most likely not. Rather it was just that the construction of the HDI did not properly consider what tradeoffs were acceptable.

Figure 1: Implicit monetary values of an extra year of life in Human Development Index



Source: Author's calculation from data provided in UNDP (2010). The fitted lines are locally smoothed (nonparametric) regressions.

The MPI and the Newsweek index also have implicit values of life, though it is hard to figure out what they are from the documentation.[22] In a recent comment on the HDI, Segura and Moya (2009) argue against imposing any tradeoff between its components, so that a country's progress in human development should be judged by the weakest (minimum) of its scaled components.

[22] In the case of the Newsweek index, scaled life expectancy gets the same weight as (say) scaled test scores for education. I will return to the MPI.

Greater clarity about the concept being measured can guide setting weights. For example, the DBI is apparently motivated by the expectation that excessive business regulation impedes investment and (hence) economic growth. Surely then a regression model of how performance in the components of the DBI has influenced these outcomes could help guide the choice of weights? Similarly, the CPIA exercise is clearly motivated by the belief that the identified attributes of country policy making matter to the goals of development aid, notably poverty reduction. Greater effort at embedding the measurement problem within a model of the relevant outcomes could help in calibrating these indices.

One of the few mashup indices that has taken seriously the problem of setting weights is the WGI. Here the weights are the estimated parameters of a statistical model, in which each of the observed indicators of governance is taken to be a linear function of an unobserved true governance measure with common parameters across countries for each indicator (Kaufmann et al., 1999; 2009, Appendix D). Under explicit distributional assumptions about this latent variable and the model's error term, the parameters can be estimated. A key identifying assumption is that the errors contained in different data sources are uncorrelated with each other—the noise in one component index is not correlated across countries with that in any other. Then the data sources that produce more highly correlated ratings can be deemed to be more informative about the latent true governance variable than sources that are weakly correlated with each other. Høyland Moene and Willumsen (2010) show how this assumption can be partially relaxed by allowing for (non-zero) correlations within certain pre-defined groups of variables. This would be important if one was to apply this method to (say) the derivation of the HDI, given that there are likely to be natural groupings of indicators; for example, the current HDI uses four variables, two of which are related to education, and can be expected to be correlated at given values of the latent concept of "human development."

However, while this approach delivers a composite index without making *ad hoc* assumptions about the weights, it is still a mashup index. The interpretation of the estimated parameters derived by this method, and (hence) the concept being measured, is far from clear. The model used to determine the weights is a statistical one, rather than economic.

Public opinion can be an important clue. A mashup index might be thought of as the first step in a public debate about what the weights should be. Stimulating such a debate would be a

valuable contribution, but there is little sign as yet that this has led to new weights. Consider, for example, the oldest of the mashup indices still in use, the HDI. Its weights were set 20 years ago, with equal weight to the (scaled) sub-indices for health, education and GDP.[23] Equality of the weights was, of course, an arbitrary judgment, and it might have been hoped that the weights would evolve in the light of the subsequent public debate. But that did not happen. The weights on the three components of the HDI (health, education and income) have not changed in 20 years, and it is hard to believe that the HDI got it right first go.[24]

Setting initial weights and revising them in the light of subsequent debate would point to the need to know the tradeoffs in the most relevant space for understanding what the weights really mean. The fact that the industry of mashup indices has often assigned weights in what can be thought of as "secondary spaces"—such as rankings or poverty measures, rather than the space of the underlying primary dimensions—does not make it easy for the debate to proceed on a well-informed basis. Indeed, given the opaqueness about the tradoffs in the primary dimensions built into most mashup indices it can be argued that users (including policy makers) may end up tacitly accepting, and acting upon, tradoffs that they would find objectionable when revealed.

Subjective questions in surveys can also offer useful clues as to the appropriate weights, although this type of data raises its own problems, such as stemming from psychological differences between respondents, including latent heterogeneity in the interpretation of the scales used in survey questions.[25] Ravallion and Lokshin (2002) discuss how subjective data on perceived economic welfare can be used to calibrate a composite index based on objective variables; the tradeoff between income and health (say) is chosen to be consistent with subjective

[23] The weights on the HDI's primary dimensions have varied over time due to (often seemingly arbitrary) changes in the bounds used for scaling the indices. However, as noted already, the weights on the HDI's core dimensions have never been explicitly identified or discussed in the HDRs. See Ravallion (2010b).

[24] In switching to a geometric mean in the 2010 HDR, the weights on the three achievement variables changed, though their logs are still equally weighted.

[25] These can stem from "frame of reference" effects, whereby a person's perception of the scales depends on the set of their own experiences and knowledge. (This is also called "differential item functioning" in the literature on educational testing; see, for example, Angoff, 1993.) In one of the few tests for such effects Beegle et al. (2009) use vignettes to anchor the scales and find that regressions using subjective welfare data are quite robust to this problem (using survey data for Tajikistan).

welfare.[26] Using survey data for Russia, the authors find that income is a highly significant predictor of subjective welfare, but that this is also influenced by health, education, employment, assets, relative income in the area of residence and expectations about future welfare.

However, for many mashup indices of development there is likely to be an important normative judgment to be made about these tradeoffs. If the index is to be accepted, then some degree of political consensus will be needed. Without that consensus, there are risks in aggregating prematurely. As Marlier and Atkinson (2010, p.292) note, "the weights are a matter for value judgments, and the adoption of a specific composite index may conceal the resolution of what is at heart a political problem."

The reality is that no consensus exists on what dimensions to include and how they should be weighted in any of the mashup indices of development in use today. And these are difficult issues. How can one contend—as the MPI does implicitly—that avoiding the death of a child is equivalent to alleviating the combined deprivations of having a dirt floor, cooking with wood, and not having a radio, TV, telephone, bike or car? Or that attaining these material conditions is equivalent to an extra year of schooling (such that someone has at least 5 years) or to not having any malnourished family member? It is very hard to say (as the MPI does implicitly) that a child's life is worth so much in terms of material goods.

And where do we draw the line in terms of what is included? In a blog comment, Duncan Green has criticized the MPI for leaving out "conflict, personal security, domestic and social violence, issues of power/empowerment" and "intra-household dynamics." Sometimes such judgments are needed in policy making at the country level. The specific country and policy context will determine what trade off is considered appropriate; any given dimension of poverty will have higher priority in some countries and for some policy problems than for others. This will typically be a political decision, though hopefully a well informed one.

But could it be that we are asking too much of a single measure of poverty to have it include things like child mortality, or schooling, or violence, as components, on top of material living standards? It is one thing to agree that consumption of market commodities is an

[26] Surveys of willingness-to-pay have also been widely used in valuation, including valuing lower risks of loss of life; in a developing-country context, see Wang and He (2010), whose results (for China) confirm intuition that the implicit value of life in developing countries built into the HDI is too low.

incomplete metric of welfare—and that for the purpose of measuring poverty one needs to account for non-market goods and services—and quite another to say that a "poverty" measure should aggregate traditional measures of (say) "human development" with command over commodities. There can be no doubt that reducing child mortality and promoting health more generally are hugely important development goals and that poverty—defined as command over (market and non-market) commodities—is an important factor in health outcomes. But does it help to have measurement efforts that risk confounding these factors in a mashup index?

How robust are the rankings given the uncertainties about data and weights?

Theory never delivers a complete specification for measurement. There is inevitably a judgment required about one or more parameters. There is also statistical imprecision about parameter estimates. Re-rankings can be generated by even very small differences in the underlying measure of interest; as Høyland et al. (2010, p.1) note, the country rankings provided by the HDI and DBI "emphasize country differences when similarity is the dominant feature."

For these reasons it is widely-recommended scientific practice to test the robustness of the derived rankings. For example, in the case of poverty measurement, where there is almost always a degree of arbitrariness about the poverty line, best practice tests the robustness of poverty comparisons to the choices made, invoking the theory of stochastic dominance.[27]

Users of prevailing mashup indices are rarely told much about the uncertainties that exist about the series chosen, the quality of the data, and their weights.[28] Few robustness tests are provided. Yet, the uncertainty about key parameters is evidently huge, and greater than other indices found in practice. It can be granted that the market prices (say) that are typically used in aggregating consumptions across commodities need not all accord with the correct shadow prices. But it is hard to accept that adding up expenditures across commodities to measure economic welfare is as problematic as valuing life, as is required by the HDI and MPI.

[27] For expositions in the standard "unidimensional" case see Atkinson (1987) and Ravallion (1994). Duclos et al. (2006) provide dominance tests for "multidimensional poverty." On ranking countries in terms of a composite index of mean income and life expectancy see Atkinson and Bourguignon (1982). Also see Anderson (2010) who applies ideas from the literature on the measurement of polarization to the task of making cross-country poverty comparisons in terms of mean income and life expectancy.

[28] An exception is the WGI, which takes seriously the imprecision in the underlying measurements of governance variables and takes account of this in its aggregation procedure, which also facilitates the construction of confidence intervals; for details see Kaufmann et al. (2010, Appendix D). The WGI is seemingly unique amongst mashup indices in this respect.

If one was to take seriously the degree of uncertainty in the data and weights, and (more generally) the functional form for aggregating across the multiple indices, one may well find that the country rankings are far from conclusive—rather dulling public interest in the mashup index. The degree of robustness to weights depends on the inter-correlations among the components. If these are perfectly correlated then nothing is gained by adding them up, and the result is entirely robust to the choice of weights. More generally, however, one expects only partial correlations.

How robust are the rankings? Some clues can be found in the literature. Slottje (1991) examines the country rankings on his own mashup index of 20 social indicators for a range of weighting methods, including averaging ranks, weights based on principal components analysis, and weights based on regression models in which a subset of the indicators were taken to be the dependent variables. Slottje's results suggest considerable sensitivity to the method used; for example, Luxembourg's rank ranges from 3 to 113 depending on the method. However, it seems that one or two of Slottje's methods might easily be ruled out as implausible.[29]

The most common method of testing robustness in this literature is to calculate the (Pearson and/or rank) correlation coefficients between alternative versions of the mashup index, such as obtained by changing the weights. The website for Doing Business reports (though with little technical detail) comparisons of the DBI's country rankings (based on the mean rank across the 10 component indicators) with rankings based on both a principal components method and "unobserved components analysis." The reported correlation coefficients with the original DBI rankings are high (0.997 and 0.982 respectively). Similarly, Kaufman et al. (2007) report results for an equally-weighted WGI (rather than the original index based on weights derived from their latent-variable model), which turns out to be highly correlated (r=0.97 or higher) with the original WGI. And Alkire et al. (2010) provide correlation coefficients between various MPIs obtained by varying the weights, with 50% weight on one of the deprivations, and 25% on each of the other two (instead of one-third on each). The correlation coefficients are all above 0.95, and they conclude that the index is "quite robust to the particular selection of weights" (p.4).[30]

[29] One of his methods seems to give perverse rankings; but even ignoring this method considerable re-ranking is evident. Luxembourg's rank ranges from 3 to 93 if one ignores the most extreme outlier method.

[30] Alkire et al (2010) also provide measures of "rank concordance," which suggest that the null hypothesis of rank independence can be rejected with 99% confidence.

However, it is not clear how much comfort one should get about robustness from even such high correlation coefficients, which can still be consistent with some sizeable re-rankings. In the case of the DBI (which provides a useful graph of the results for the alternative methods), the largest change appears to be a country (un-named) whose rank falls from about 50 using the ordinary DBI to 80 using the unobserved components ranking.

In the case of the CPIA, the country rankings do not play any role in the World Bank's aid allocations, which are based on the aforementioned "governance-heavy" index based on the CPIA. This re-weighted index turns out to be highly correlated with the original (equally-weighted) index; the correlation coefficient is 0.96 using the 2009 CPIA.[31] This is not surprising given that the components of the CPIA are highly correlated amongst themselves. Across the 77 countries receiving concessional loans under IDA, the correlation coefficients with the CPIA are 0.86 for its "economic management" component, 0.87 for "structural policies," 0.91 for "social inclusion/equity," and 0.90 for "public sector management." Given these high correlations, the index is affected little by changes in its weights.

The fact that the ordinary CPIA and this re-weighted index have a correlation coefficient of 0.96 might be taken to suggest that extra weight on governance is largely irrelevant. However, that reasoning ignores the fact that, in attempting to reward good policies (particularly on governance), the IDA allocation per capita is highly elastic—an elasticity of five—with respect to the index (International Development Association, 2008, Annex 1). Then changes in the weights will matter to aid allocations. This is evident if one compares the actual aid allocations under IDA with those implied by the ordinary (equally-weighted) index based on the CPIA. To make the comparison (approximately) budget neutral I have rescaled the equally-weighted index to have the same mean as the actual index used by IDA. Then I find that the implied proportionate changes in IDA allocation in switching from the equally-weighted CPIA-based index to the governance-heavy index range from 0.68 to 2.49. Of the 77 countries receiving concessional loans under IDA, I estimate that 16 would have seen their allocation increase by at least 20% with the higher weight on governance, while 15 countries would see it fall by 20% or

[31] In calculating the re-weighted index I used a weight of 0.74 on governance and 0.26 on the mean of the other three components; the relative weights are the same as used for IDA allocations, though the absolute weights differ slightly given that another variable enters into the allocations, as noted above.

more. Despite the high correlations, it is clear that changing the weights makes a sizable difference to aid allocations.

Data and methodological revisions also provide a clue to the robustness of mashup indices. An independent evaluation of the DBI by the World Bank (2008) pointed to a number of concerns about the robustness of country rankings to data revisions. The evaluation found 2,200 changes to the original data posted in 2007; the data revisions changed the country rankings by 10 or more for 48 countries. Wolff et al (2010) use data revisions to measure the imprecision in the HDI, and find standard deviations that vary from 0.03 (for the United States) to 0.11 for Niger (recall that the HDI is scaled to the (0,1) interval). Poorer countries tend to have less accurate HDIs.

In the case of mashup indices that use expert assessments, such as the CPIA, we can learn about robustness by comparing the assessments of different experts. The same CPIA questionnaire administered to the World Bank's country experts was also completed by experts at the African Development Bank (though only for Africa of course). Kaufmann and Kraay (2008) compared the two and found many notably differences in the CPIA ratings for 2005. The overall correlation coefficient in the two institution's scores on governance across the countries of Africa was significantly positive, with a correlation coefficient of 0.67, but still suggests a good deal less than full agreement. Of course, the source of these differences is unclear. Experts may disagree on the facts about a country, or they may disagree about how those facts are to be weighted in forming the various sub-indices that go into the CPIA.

I repeated this test for the 2009 CPIA ratings of governance by both institutions, and found that the correlation has risen to 0.87. The correlations are similar for other CPIA components: for economic management the coefficient is 0.88, for structural policies it is 0.85, while it is 0.87 for social inclusion/equity.[32] The correlation coefficient between the overall CPIA indices is 0.94. While their expert assessments cannot be considered independent, these correlations point to a high level of agreement, with signs that this has risen over time. However, as already noted, the aid allocations based on these indices may well be sensitive to even small differences, depending on the allocation formulae.

[32] These calculations use the 2009 CPIA ratings available at relevant World Bank and African Development Bank (ADB) web sites. There are 39 countries with CPIA ratings from both institutions.

In 2010, the Human Development Report introduced a number of changes to the data and methods of the HDI (UNDP, 2010). Ravallion (2010) shows that these changes led to a marked reduction in the implicit monetary valuation of extra longevity, notably in low and middle-income countries; the whole schedule in Figure 1 was noticeably higher using the prior HDI method (though even then some observers felt that the implied valuations of life were too low). The change in aggregation method generated large downward revisions in the HDIs for Sub-Saharan Africa. The reasons for the data and methodological changes are not entirely clear from the report, though the main reason given is the desire to relax the perfect substitution property of the old HDI, whereby the MRS was constant between the sub-components.

Ravallion (2010) provides an alternative HDI, based on Chakravarty's (2003) generalization of the HDI. This alternative index also allows imperfect substitution but has advantages over the new HDI proposed by UNDP (2010); in particular, the valuations on longevity appear to be more plausible, and show only a mild income gradient. Figure 1 also gives the valuations of longevity implied by this alternative index. While the two HDIs are highly correlated (r=0.980), there are many large changes. Zimbabwe's index rises by over 300%, from the lowest value (by far) of 0.14 based on the UNDP's (2010) index to 0.45 using the alternative HDI; it also rises relatively, to be the 12th lowest—reflecting the fact that the additivity property of the Chakravarty index allows it to give a higher reward for Zimbabwe's relatively good schooling attainment. The largest decrease in the HDI is that for New Zealand, for which the index falls by 0.094 and the ranking falls from third place to 18th. The largest increase in ranking when switching to the Chakravarty index is for Qatar, which rises from the 38 highest using the 2010 HDI to third place.

Confidence intervals (CIs) provide a basis for assessing robustness. This is not common practice in this literature, though an exception is the WGI, for which the econometric method used to estimate the weights readily delivers standard errors (Kaufman et al., 1999). Høyland et al. (2010) apply a version of the WGI method to both the HDI and DBI to test the robustness of their country rankings.[33] They find wide confidence intervals for both the HDI and DBI (both using data for 2008), indicating that the rankings can be highly sensitive, though less so at the extremes. For example, Singapore, New Zealand, the US and Hong Kong are deemed by

[33] They use a Bayesian estimation method, also taking account of the ordinal nature of some of the data.

Høyland et al. to be "almost surely" in the top 10 of the DBI, while Congo, Zimbabwe, Chad and the Central African Republic are almost surely among the 10 countries doing worst. However, most rankings in the middle 80% look far more uncertain. Høyland et al. (2010, p. 15) conclude:

> "In contrast to the key message of the precise ranking published in the Doing Business report, it is clear that the index does not do a very good job in distinguishing between most of the regulatory environments in the world. While the rankings, after taking uncertainty into account, clearly distinguish the best economies from the worst, it does not distinguish particularly well between the economies that are somewhere in between."

Turning to the HDI, Høyland et al. find that no country has more than a 75% chance of being in the top 10 in terms of this composite index, though we can have more confidence about which countries have very low HDIs. Similarly to the DBI, there is great uncertainty about the middle rankings. For example, Georgia has a DBI rank of 18, but Høyland et al. find that the 95% CI is that the true ranking lies between 11 and 59. To give two more examples, Saudi Arabia has a DBI rank of 23 but a 95% CI of (12, 63), while for Mauritius, with a DBI rank of 27, the CI estimated by Høyland et al. is (16, 77).

In the light of their findings, Høyland et al. argue that it would be more defensible for these composite indices to try to identify a few reasonably robust country groupings than these seemingly precise but actually rather uncertain country rankings. Of course, there will always be a degree of arbitrariness about such groupings; for example, the 2010 edition of the HDR uses quartiles. However, Høyland et al. provide defensible country groupings for the HDI (and DBI) for various "certainty thresholds," given by one's desired confidence that there is a difference between the top and bottom ranked country within a given group.

The EPI has been subjected to numerous sensitivity tests, reported in Saisana and Saltelli (2010). They find that the rankings for 60 of the 163 countries "…depend strongly on the original methodological assumptions made in developing the Index and any inference on those countries should be formulated with great caution" (p.3). For the other 103 countries, the ranking was considered reasonably robust, although this only means that the actual EPI rank lies within a confidence interval that could span up to 20 positions in the country ranking.[34]

Probing some of the data provided on the websites for recent mashup indices also helps give us an idea of their sensitivity to different weights. For example, I find that Finland's ranking

[34] Also see the results on the EPI reported in Foster et al. (2009).

as number 1 in Newsweek's index falls to 17 if I put all the weight on health; Australia's rank at number 4 falls to 13 if one put all the weight on education. In exploring the website for Newsweek's mashup, the most dramatic impact of re-weighting appears to be for China; if one puts all the weight on "economic dynamism" China's rank rises from 66 to 13.

None of their websites make it easy for users to properly assess the sensitivity of these mashup indices to changing weights. Yet it would be relatively easy to program the required flexibility into the current web sites, so users can customize the index with their preferred weights, to see what difference it makes. The only example I know of to date is the OECD's Social Institutions and Gender Index. Their interactive website, "My Gender Index," allows users to vary the composition and weights of the index, and immediately gives the corresponding country rankings and maps them. There are also some useful graphical tools for assessing robustness from the work of Foster et al. (2009). A careful assessment of robustness using such tools would be a more open approach than encouraging users to think that the data have been aggregated in the one uniquely optimal way.

Few of the mashups of development data have said much about data quality, including international comparability. Data constraints are often mentioned, but most of the time the mashups take their data as given with little or no critical attention to the problems; the data often come from others who can be blamed for its inadequacies.[35] Under certain circumstances, forming a mashup index may actually help reduce data concerns, notably when averaging across indicators reduces overall errors. This may have bearing on the choice of indicators, though one finds little sign in the documentation on past mashup indices that this has been considered.

Possibly more worrying than the lack of attention to data quality in existing mashups is how little is done to expose and address the problems in pre-existing data series. The rapid growth in mashup indices will hopefully come with greater attention to these problems, though that may well be little more than hope unless prevailing practices change on the part of mashup producers; greater critical scrutiny and skepticism from mashup consumers would help.

A cavalier approach to data issues appears at times to come hand-in-hand with immodesty in the claims made about new knowledge generated by simply aggregating pre-

[35] An exception is the DBI, which relies on primary data collected by the team.

existing data. "Important new insights" are claimed about (for example) the causes of poverty and how best to fight it even though there has been no net addition to the stock of data—just a re-packaging of what we already had—and no sound basis is evident for attributing causation.[36]

How is the index useful for development policy? If we agreed that the index provides an adequate characterization of some development goal, and that its embodied tradeoffs are acceptable, what would we do with it?

An important role served by mashup indices can be to provide an easily administered antidote to overly narrow conceptualizations of development goals. Putting aside the straw-man argument that GDP is seen as the sole measure of welfare, the HDI has helped sensitize many people to the importance of aspects of human welfare that are not likely to be captured well by command over market goods. This can provide a useful re-balancing when policy discussions appear to put too little weight on factors such as access to public services in determining undeniably important aspects of human welfare such as health (Anand and Ravallion, 1993).

Does this translate into better development policies? It has been argued that country comparisons of a mashup index can influence public action in those countries that are ranked low. This has been claimed by proponents of both the HDI and DBI. In the context of the HDI, there is an interesting discussion of this point in Srinivasan (1994, p.241), who argues that:

> "… there is no evidence that HDR's have led countries to rethink their policies, nor is there any convincing reason to expect it to happen. It was widely known, long before the first HDR in 1990, that in spite of her low per capita real income Sri Lanka's achievements in life expectancy and literacy were outstanding, in comparison not only with neighbors, but also with countries (developed and developing) with substantially higher per capita in-comes. This knowledge did not demonstrably lead other countries to learn from Sri Lanka's experience. An even more telling example is that of the Indian state of Kerala with its substantially lower rates of infant and child mortality and higher rates of literacy in comparison to other states, including Punjab with more than twice Kerala's real domestic product per capita. Yet such disparities in performance within the same country have not led to significant policy changes in the lagging regions. Surely socio-economic-political processes, rather than low levels of income and lack of knowledge about the feasibility of achieving substantial improvements, precluded the policy changes needed to bring about improvement."

[36] For example, in the press release (PR) for the MPI, one of the authors is quoted as saying that "The MPI is like a high resolution lens which reveals a vivid spectrum of challenges facing the poorest households." The PR does not point out that the MPI relies entirely on existing publicly available data. The contribution of the MPI is to mashup these data.

On thinking about this issue 16 years after Srinivasan was writing, it appears that a degree of cross-country learning has emerged among developing countries. However, one can broadly agree with Srinivasan that it is not comparisons of country rankings in terms of some mashup index that have been the main driving force in that learning process; rather it is the comparison of experiences with specific policies, and the process of adapting those policies to new settings. The learning process about anti-poverty policies provides examples, of which the most prominent in recent times is the set of policies known as Conditional Cash Transfers, where a now famous program in Mexico, *PROGRESA* (now called *Oportunidades*), has been cloned or adapted to many other countries.[37] To the extent that a country government learns about seemingly successful policy experiences elsewhere via seeing its low ranking in some mashup index, the latter will have contributed to better policies for fighting poverty. However, it does not appear likely that this is how the learning typically happens, which appears to be more directly focused on the space of policies than country rankings in terms of the mashup index.

If a country was keen to improve its ranking and the index is sufficiently transparent about how it was constructed, it should be clear what the country's government needs to do: it should focus on the specific components of the index that it is doing poorly on. This is what Høyland et al. (2010) dub "rank-seeking behavior." It has been claimed that the DBI (or at least some specific components, notably business entry indicators) have stimulated policy reforms to improve country rankings based on the index.[38] Although the attribution to the DBI would seem difficult to establish. It has been argued that the mashup index plays a key role in promoting such reforms. The Doing Business website argues that a single ranking of countries has the advantage that "it is easily understood by politicians, journalists and development experts and therefore creates pressure to reform." Of course, the reform response will then focus on those components of the index that rank low and are easily changed. Anecdotally, a Cabinet Minister in a developing country (that will remain nameless to preserve confidentiality) once told me that he had been instructed by his President to do something quickly about the country's low ranking in

[37] For further discussion see Fishbein and Schady (2009). The Mexico program had antecedents in similar types of policies found elsewhere, including Bangladesh's Food for Education Program and the means-tested school bursary programs found in some developed countries.

[38] A page on the Doing Business web site claims "26 reforms have been inspired or influenced by the Doing Business project."

the DBI.[39] The Minister picked the key indicators, and by a few relatively simple legislative steps, was able to improve the country's ranking. But these indicators were only *de jure* policy intentions, with potentially little bearing on actual policy implementation at the firm level. Deeper characteristics of the business and investment climate in the country did not apparently change in any fundamental way, and the Minister felt that there was no genuine impact on the country's development.

Nor should it be presumed that efforts to improve a county's ranking by manipulating the few proxies for poor performance that happened to get selected for the mashup are costless. Targeting reform efforts on a few partial indicators, which on their own may bring little gain, can have an opportunity cost. This has been an issue with DBI. Arrunada (2007) argues that an exclusive focus on (for example) simplifying the procedures for business start-ups risks distorting policy by not putting any weight on the benefits (to firms and the public at large) derived from formal registration procedures.

There are also applications of mashup indices, along with other composite indices, as explanatory variables in policy-relevant models for outcomes of interest. For example, the Doing Business indices have been widely used in a (large) academic literature as explanatory variables for (*inter alia*) productivity, entrepreneurship and corruption.[40] Such applications are potentially important, although arguably it is the component series that should be the regressors not the composite index, letting the regression coefficients set the weights, appropriate to the specific application.[41] In this case the dependent variable provides the relevant basis for setting weights, and the mashup index can be discarded.

It is not obvious how useful an aggregate (country-level) mashup index is for policy making in a specific country. Development policy making has increasingly turned instead to micro data on households, firms and facilities. These are data on both the outcomes of interest and instrumentally important factors, including exposure to policy actions. Such micro data invariably reveal heterogeneity in outcomes and policies within countries. As Hallward-Driemeier, Khun-Jush and Pritchett (2010) argue, the *de jure* representation of policies at

[39] Høyland et al. (2010) give other examples of such rank-seeking behavior.
[40] A useful compendium of research using these data can be found on the Doing Business web site. Also see Djankov's (2009) survey.
[41] See Lubotsky and Wittenberg (2006) for a formal exposition of this argument.

country level (such as used in the DBI) may actually be quite deceptive about *de facto* policy impacts on the ground. *De jure* rules may have little relationship with the incentives and constraints actually facing economic agents. Indeed, Hallward-Driewmer et al. find virtually no correlation in Africa between country-level policies and policy actions reported in micro enterprise data; the within-country variation in the latter exceeds the between-country variation in *de jure* rules. This reflects the potential for idiosyncratic deals by firms to get around rules.

The (domestic and international) policy relevance of any composite index of development data is also questionable in the absence of any "contextuality"—the many conditions that define the relevant constraints on country performance. It is not credible that any one of these indices could be considered a sufficient statistic for country performance even with regard to the development outcome being measured. Very poor countries invariably fare poorly in the rankings by the various indices discussed above. However, these indices tell us nothing about how we should judge the performance of these countries, given the constraints they face. We may well rank them very differently if we took account of the country's stage of economic development. Such conditional comparisons raise their own concerns that need to be taken seriously, as discussed in Ravallion (2005). However, without greater effort to allow for the circumstances and history of a country it is not clear what we learn from the index. The greater use of benchmarking and time series comparisons will help here, though we also have to be aware of the fact that differing initial conditions at the country level can have lasting effects on a country's development path.

Policy applications also call for greater transparency about the tradeoffs built into the index. Consider a simple characterization of the problem of allocating public resources across a set of indicators that have been aggregated into a composite index. The policy maker has a set of policy instruments available for improving the index. Let us also assume that these policy instruments have known costs that can be mapped one-to-one to the underlying indicators. A policy maker deciding how best to improve the composite index by shifting resources between any two components should compare their MRS in the composite index with the relative marginal costs of the corresponding policy instruments. And the optimal allocation of a given

budget will equate the MRS with the ratio of those marginal costs.[42] Yet, as we have seen, many existing mashup indices have said little or nothing about those tradeoffs. Unless the mashup index considers, and reveals, its MRSs across components, or its marginal weights, it will be impossible to assess whether it is acceptable as a characterization of the development objective, and impossible to advise how policy can best be aligned with that objective.

If one un-packs the aggregate index, a potential application is in allocating central funds across geographic areas—the "targeting problem." Here the value-added of the mashup aggregation becomes questionable if its components can be mapped (at least roughly) to policy instruments; indeed, that is sometimes why the data were collected in the first place. Then the obvious first step when given a mashup index is to un-pack it. The actionable things based on such data are not typically found in the composite itself but in its components. Thankfully, many of the mashup indices found in practice can be readily un-packed, though it remains unclear what policy purpose was served by adding them up in the first place.

This point is illustrated well by proposals to use "multidimensional poverty" indices for targeting. The MPI is intended to inform policy making. Alkire and Santos (2010b, p.7) argue that:

> "The MPI goes beyond previous international measures of poverty to identify the poorest people and aspects in which they are deprived. Such information is vital to allocate resources where they are likely to be most effective."

But is it the MPI or its components that matter for this purpose? Following Alkire and Foster (2007), the MPI has a neat decomposability; we can reverse the mashup aggregation. This is useful, for only then will we have any idea how to go about addressing the poverty problem in that specific setting. Should we be focusing on public spending to promote income growth or better health and education services? Consider the following stylized example (simplifying the MPI for expository purposes). Suppose that there are two dimensions of welfare, "income" and "access to services." Assume that an "income-poor" but "services-rich" household attaches a high value to extra income but a low value to extra services, while the opposite holds for an

[42] This statement requires certain restrictions on the curvatures of the relevant functions, which I will ignore for the purpose of this discussion.

"income-rich" but "services-poor" household.[43] There are two policy instruments, a transfer payment and service provision. The economy is divided into geographic areas and a given area gets either the service or the transfer. We then calculate a composite index like the MPI based on survey data on incomes and access to services. There is bound to be a positive correlation between average income and service provision, but (nonetheless) some places have high income poverty but adequate services, while others have low income poverty but poor services. The policy maker then decides whether each area gets the transfer or the service. Plainly, the policy maker should not be using the aggregate MPI for this purpose, for then some income-poor but service-rich households will get even better services, while some income-rich but service poor households will get the transfer. The total impact on (multidimensional) poverty would be lower if one based the allocation on the MPI rather than the separate poverty measures—one for incomes and one for access to services. It is not the aggregate mashup index that we need for this purpose but its components. Indeed, for such policy applications we do not need the mashup index.

Conclusions

The lesson to be drawn from all this is not to abandon mashup indicies. Composite indices derived from development-data mashups are often trying to attach a number to an important, but unobserved, concept, for which prevailing theories and measurement practices offer little guidance. And there are clear attractions to finding a way of collapsing a (potentially) large number of dimensions into one. Rather the main lessons are (first) that the current enthusiasm for new mashup indices needs to be balanced by clearer warnings for, and more critical scrutiny from, users, and (second) that some past mashup indices do not stand up well to such scrutiny.

While there is invariably a gap between the theoretical ideal and practical measurement, for past mashup indices the gap is huge. Greater clarity is needed on what exactly is being measured. And more attention needs to be given to the tradeoffs embodied in the index. In most cases the tradeoff is not even identified in the most relevant space for users to judge, and in cases where it can be derived from the data available it has been found to be questionable—implying,

[43] Sufficient conditions are that there is declining marginal utility to both income and services and that the marginal utility of income (services) is non-decreasing in services (income).

for example, unacceptably low valuations of life in poor countries. There is a peculiar inconsistency in the literature on mashup indices whereby prices are regarded as an unreliable guide to tradeoffs, and are largely ignored, while the actual weights being assumed in lieu of prices are often not made explicit in the same space as prices. Thus we have no basis for believing that the weights being used are any better than market prices, when available. Nor do we have any basis for believing that the weights bear any resemblance to defensible shadow prices. Aggregating under such conditions risks stifling, rather than promoting, open debate about what tradeoffs are in fact acceptable, when such tradeoffs need to be set.

Mashup producers need to be more humble about their products. The rhetoric of these indices is often in marked tension with the reality. Not all are as ambitious as Newsweek's effort to find the "World's Best Countries" using a mashup of mashups. But exaggerated claims are not uncommon even in the more academic efforts. One is struck, for example, that the "multidimensional poverty indices" proposed to date actually embrace far fewer dimensions of welfare than commonly-used measures based on consumption at household level. Arguably, the seeming precision of these mashup indices and their implied country rankings (so closely watched by the media) is more an illusion than real given the considerable uncertainties about the data and how they should be aggregated. As some commentators have suggested, it would be more defensible to try to identify broad country groupings rather than precise rankings of individual countries.

The uncertainty about the components and their weights is not adequately acknowledged by mashup producers, and users are given little guidance to the robustness of the resulting country rankings. Today's technologies permit greater openness about the sensitivity of country rankings to choices made about a mashup index's (many) moving parts. For non-market goods it appears to be highly implausible that the weights would be constant across everyone in a given country, let alone across all the countries (and peoples) of the world. Knowing nothing else about their design, this fact alone must make one skeptical of past mashup indices.

Policy relevance is often claimed, but is rarely so evident on close inspection. It is unclear what can be concluded about "country performance" toward agreed development goals in the absence of an allowance for the (country-specific) contextual factors that constrain that performance. (The words "performance" and "impact" are used too loosely in the mashup

industry, though this is also true in some other areas of policy discourse.) There are also potentially important "targeting applications," though here policy-makers might be better advised to use the component measures appropriate to each policy instrument rather than the mashup index.

With greater attention to such issues, thoughtful users of these increasingly popular indices of development will be better informed, and better able to judge the merits of the index. Some of the mashup indices in recent times have contributed to our knowledge about important development issues, though arguably much of this was achieved by the primary data collection efforts rather than the mashup *per se*. In the absence of more convincing efforts to address the concerns raised by this paper, we should not presume that mashups of pre-existing development data have taught us something we did not know—adding explanation, understanding or insight where there was none before. That is not what happened when the mashup index was formed. Rather, it took things we already knew and re-packaged them, and too often in a way that will be opaque to many users, and yet contentious if those users understood what went into the mashup.

Arguably, mashup indices exist because theory and rigorous empirics have not given enough attention to the full range of measurement problems faced in assessing development outcomes. The lessons for measurement from prevailing economic theories only take us so far in addressing the real concerns that practitioners (including policy makers) have about current measures. A mashup index is unlikely to be a very satisfactory response to those concerns. Theory needs to catch up. It also needs to be recognized that the theoretical perspectives relevant to measurement practice are not just found in economics, but also embrace the political, social and psychological sciences.

Thankfully, progress in development does not need to wait for that catch up to happen. A composite index is not essential for many of the purposes of evidence-based development policy-making. Recognizing the multidimensionality of development goals does not imply that we should be aggregating fundamentally different things in opaque and often questionable ways. Rather it is about explicitly recognizing that there are important aspects of development that cannot be captured in a single index.

References

African Development Bank, 2007, *Investing in Africa's Future: The ADB in the 21st Century.* Tunis: African Development Bank.

Alkire, Sabina and James Foster, 2007, "Counting and Multidimensional Poverty Measurement," Oxford Poverty and Human Development Initiative, Working Paper 7, University of Oxford.

Alkire, Sabina and Maria Emma Santos, 2010a, "Acute Multidimensional Poverty: A New Index for Developing Countries," Oxford Poverty and Human Development Initiative, Working Paper 38, University of Oxford.

_____ and _____, 2010b, "Multidimensional Poverty Index," Research Brief, Oxford Poverty and Human Development Initiative, University of Oxford.

Alkire, Sabina, Maria Emma Santos, Suman Seth and Gaston Yalonetzky, 2010, "Is the Multidimensional Poverty Index Robust to Different Weights?" mimeo, Oxford Poverty and Human Development Initiative, University of Oxford.

Anand, Sudhir and Martin Ravallion, 1993, "Human Development in Poor Countries: On the Role of Private Incomes and Public Services," *Journal of Economic Perspectives* 7: 133-150.

Anand, Sudhir and Amartya Sen, 2000, "The Income Component of the Human Development Index," *Journal of Human Development* 1(1): 83-106.

Anderson, Gordon, 2010, "Polarization of the Poor: Multivariate Relative Poverty Measurement sans Frontiers," *Review of Income and Wealth* 56: 84-101.

Angoff, William H., 1993, "Perspectives on Differential Item Functioning Methodology." in Paul Holland and Howard Wainer (eds) *Differential Item Functioning*, Lawrence Erlbaum Associates.

Arrunada, Benito, 2007, "Pitfalls to Avoid when Measuring Institutions: Is Doing Business Damaging Business?", *Journal of Comparative Economics* 35: 729-747.

Atkinson, Anthony B., 1987, "On the Measurement of Poverty," *Econometrica* 55: 749-64.

Atkinson, Anthony B., and Francois Bourguignon, 1982, "The Comparison of Multi Dimensional Distributions of Economic Status," *Review of Economic Studies* 49: 183-201.

Beegle, Kathleen, Kristen Himelein and Martin Ravallion, 2009, "Testing for Frame-of-Reference Bias in Subjective Welfare," Policy Research Working Paper 4904, World Bank, Washington DC.

Blackorby, C. and Donaldson, D., 1987, "Welfare Ratios and Distributionally Sensitive Cost-Benefit Analysis," *Journal of Public Economics* 34: 265–90.

Bourguignon, Francois and Satya Chakravarty, 2003, "The Measurement of Multidimensional Poverty," *Journal of Economic Inequality* 1: 25-49.

Browning, Martin, 1992, "Children and Household Economic Behavior," *Journal of Economic Literature* 30(3): 1434-75.

Chakravarty, Satya R., 2003, "A Generalized Human Development Index," *Review of Development Economics* 7(1): 99-114.

Chakravarty, Satya, J. Deutsch and Jacques Silber, 2008, "On the Watts Multidimensional Poverty Index and its Decomposition," *World Development* 36(6): 1067-78.

Chen, Shaohua and Martin Ravallion, 2010, "The Developing World is Poorer than we Thought, but no Less Successful in the Fight Against Poverty," *Quarterly Journal of Economics* 125(4): 1577–1625.

Djankov, Simeon, 2009, "The Regulation of Entry: A Survey," *World Bank Research Observer* 24(2): 183-203.

Djankov, Simeon, Rafael La Porta, Florencio Lopez-de-Silanes and Andrei Shleifer, 2002, "The Regulation of Entry," *Quarterly Journal of Economics* 117(1): 1-37.

Duclos, Jean-Yves, David Sahn and Stephen Younger, 2006, "Robust Multidimensional Poverty Comparisons," *Economic Journal* 116: 943-68.

Fiszbein, Ariel and Norbert Schady, 2009, *Conditional Cash Transfers for Attacking Present and Future Poverty,* World Bank Policy Research Report, World Bank, Washington DC.

Foster, James, J. Greer, and Erik Thorbecke, 1984, "A Class of Decomposable Poverty Measures", *Econometrica* 52: 761-765.

Foster, James, Mark McGillivray and Suman Seth, 2009, "Rank Robustness of Composite Indices," OPHI Working Paper 29, Oxford University.

Gwartney, James and Robert Lawson, 2009, *Economic Freedom of the World: 2009 Annual Report*, Fraser Institute, Vancover, Canada.

Hallward-Driemeier, Mary, Gita Khun-Jush and Lant Pritchett, 2010, "Deals versus Rules: Policy Implementation Uncertainty and Why Firms Hate It," Policy Research Working Paper 5321, World Bank, Washington DC.

Høyland, Bjørn, Kalle Ove Moene and Fredrik Willumsen, 2010, "The Tyranny of International Index Rankings," mimeo, Department of Economics, University of Oslo.

International Development Association, 2008, *IDA15. Report of the Executive Directors of the International Development Association to the Board of Governors.* Washington DC: World Bank.

Jones, Charles and Peter Klenow. 2010, "Beyond GDP? Welfare Across Countries and Time." Mimeo, Stanford University.

Kaufmann, Daniel and Aart Kraay, 2008, "Governance Indicators: Where Are We, Where Should We Be Going?" *World Bank Research Observer* 23(1): 1-30.

Kaufmann, Daniel, Aart Kraay and Pablo Zoido-Lobatón, 1999, "Aggregating Governance Indicators." World Bank Policy Research Working Paper No. 2195, Washington, D.C.

Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi, 2007, "The Worldwide Governance Indicators Project: Answering the Critics," Policy Research Working Paper 4149, World Bank, Washington, D.C.

_____, _____ and _____, 2009, "Governance Matters VIII. Aggregate and Individual Governance Indicators 1996-2008," Policy Research Working Paper 4978, World Bank, Washington DC.

Kelley, A.C., 1991, "The Human Development Index: 'Handle with Care,'" *Population and Development Review* 17(2): 315-324.

Lubotsky, Darren and Martin Wittenberg, 2006, "Interpretation of Regressions with Multiple Proxies," *Review of Economics and Statistics* 88(3):549-62.

Marlier, Eric and Anthony B. Atkinson, 2010, "Indicators of Poverty and Social Exclusion in a Global Context," *Journal of Policy Analysis and Management* 29(2): 285-304.

Morris, M. D., 1979, *Measuring the Condition of the World's Poor: The Physical Quality of Life Index,* Washington, D.C.: Overseas Development Council.

Pollak, Robert and Wales, T., 1979, "Welfare Comparison and Equivalence Scale," *American Economic Review* 69: 216–21.

Ravallion, Martin, 1994, *Poverty Comparisons,* Chur, Swtizerland: Harwood Academic Press.

_____, 1997, "Good and Bad Growth: The Human Development Reports," *World Development*, 25(5): 631-638.

_____, 2005, "On Measuring Aggregate 'Social Efficiency'," *Economic Development and Cultural Change*, 53 (2): 273-92.

_____, 2008, "Poverty Lines," in *The New Palgrave Dictionary of Economics*, 2nd Edition, Larry Blume and Steven Durlauf (eds) London: Palgrave Macmillan.

_____, 2010a, "On Multidimensional Indices of Poverty." *Journal of Economic Inequality*, forthcoming.

_____, 2010b, "Troubling Tradeoffs in the Human Development Index." Policy Research Working Paper 5484, World Bank, Washington DC.

Saisana, Michaela and Andrea Saltelli, 2010. *Uncertainty and Sensitivity Analysis of the 2010 Environmental Performance Index.* Luxembourg: Office of Official Publications of the European Communities.

Samuelson, Paul A., 1983, *Foundations of Economic Analysis. Enlarged Edition.* Harvard: Harvard University Press.

Segura, Sebastian Lozano and Ester Gutierrez Moya, 2009, "Human Development Index: A Non-Compensatory Assessment," *Cuadernos de Economia* 28: 223-35.

Sen, Amartya, 1985, *Commodities and Capabilities*. Amsterdam: North-Holland.

_____, 1999, *Development as Freedom*, New York: Alfred Knoff.

Slesnick, Daniel, 1998, "Empirical Approaches to the Measurement of Welfare," *Journal of Economic Literature* 36(4): 2109-2165.

Slottje, Daniel J., 1991, "Measuring the Quality of Life Across Countries," *Review of Economics and Statistics* 73(4): 684-693.

Srinivasan, T.N. 1994, "Human Development: A New Paradigm or Reinvention of the Wheel?" *American Economic Review, Papers and Proceedings*, 84(2): 238-249.

Stiglitz, Joseph, Amartya Sen and J.P. Fitoussi JP, 2009, *Report by the Commission on the Measurement of Economic Performance and Social Progress*, www.stiglitz-sen-fitoussi.fr.

Tsui, Kai-Yuen, 2002, "Multidimensional Poverty Indices," *Social Choice and Welfare* 19: 69-93.

United Nations Development Programme (Various Years) *Human Development Report*, New York: Oxford University Press or Palgrave Macmillan for the UNDP.

Wang, Hua and Jie He, 2010, "The Value of Statistical Life: A Contingent Investigation in China," Policy Research Working Paper 5421, World Bank, Washington DC.

Watts, Harold W., 1968, "An Economic Definition of Poverty," in Daniel P. Moynihan (ed.), *On Understanding Poverty*, New York, Basic Books.

Wolff, Hendrik, Howard Chong and Maximilian Auffhammer, 2010, "Classification, Detection and Consequences of Data Error: Evidence from the Human Development Index," NBER Working Paper 16572, National Bureau of Economic Research.

World Bank, 2008, *Doing Business: An Independent Evaluation.* Washington DC: Independent Evaluation Group, World Bank.

World Bank, 2009, *World Development Indicators*, Washington DC: World Bank.