‘Valuing development’: Could approaches to measuring outcomes in health help make development more accountable?

Claire Melamed, Nancy Devlin & John Appleby

March 2012

Acknowledgements

ODI gratefully acknowledges the support of the Bill and Melinda Gates Foundation in the production of this report.

Overseas Development Institute
111 Westminster Bridge Road
London SE1 7JD, UK

Tel: +44 (0)20 7922 0300
Fax: +44 (0)20 7922 0399
www.odi.org.uk

Disclaimer: The views presented in this paper are those of the author(s) and do not necessarily represent the views of ODI or our partners.

Contents

1 The problem

Consider this quote, from a leading professional in his field:

> *"The ultimate measure by which to judge the quality of an... effort is whether it helps ... families as they see it. Anything done ... that does not help a ... family is, by definition, waste, whether or not the professions and their associations traditionally hallow it."*

Most development professionals would surely agree with this sentiment. However, the quote is not from an academic or an aid practitioner, but from Donald Berwick, outgoing Administrator of the USA's Medicaid system. Berwick is talking about health care. But the sentiment could, and should, apply to development as well. Whether over the short term, as in a cash transfer programme, or over the long term, as in the development of transport infrastructure, development should be about making lives better for people as they themselves understand it.

But does what outsiders do in the name of development actually do the very best it can to accomplish this? It is a hard question to answer, since poor people are rarely asked about their priorities. But there are worrying signs that it does not. For example, when asked, people usually put jobs pretty high on the list of desirable outcomes – but jobs have been very low on donor priorities for many years. Improved infrastructure is usually also close to the top of people's wish-lists – again, it has been near the bottom of the donor list. Likewise people tend to rate their own physical security very highly, but this issue has not been of much interest to the donor community until recently.

The answer to the question might, then, be no. But the question itself has become more immediate since the development business is being asked to get more serious about results and evaluation. The urgings of politicians like Andrew Mitchell in the UK (Mitchell, 2011) and Raj Shah in the USA (Shah, 2011) that aid spending must provide good value for money and media clamour are focusing minds and research energy on how aid is spent, on what, and on how we know if the interventions that aid finances are actually working.

This is an opportune time to think about the outcomes that 'development' is trying to achieve and who gets to decide what they are.

If the drive for results and value for money is to deliver improved benefits for poor people, and better evidence for politicians trying to justify aid spending to taxpayers, we need an improved metric that allows us to *measure* outcomes in a standard way, and identify which – of the possible outcomes that aid spending could achieve – poor people would *value* the most. To be effective, this metric needs to be communicated in a way that is easy to use for decision makers in the development business.

This information could feed into the allocation of resources and planning of interventions, to ensure that the 'value' in value for money is informed by what people themselves value. It could also be used in evaluations, to assess the extent to which interventions are delivering the changes of most

value to poor people, and in ongoing monitoring to assess whether value is being delivered over time.

Development is a complex business and any such measure would inevitably involve huge simplifications of reality. The persistence of income-based measures of development outcomes shows how the need for simple indicators for policy makers can sometimes outweigh the need to represent this complexity in an adequate way***. The challenge, and the research question, is to establish whether we can devise a way of measuring outcomes, based on poor people's own values, which strikes the right balance between a reasonable representation of reality and usability by policy makers. This report proposes one approach to answering this question.***

2 The political economy of development outcomes

The question of how to define outcomes, measure progress, and trade off different possible uses of resources is of course the day to day problem of politics in every country. Every public body faces the challenge – from the English National Health Service (NHS), to the Ghanaian Ministry of Education, to the administrators of Brazil's cash transfer programme.

It follows that the question of defining outcomes based on the views of the people affected should not be seen as a problem unique to development but to public policy more generally. It is the norm in most countries that the poorest are the least able to influence policy to their benefit – therefore, the lack of fit between priorities for development spending and poor people's own priorities should not be surprising.

If all development aid were given directly and without any conditions upon recipient governments, then the political economy of development, like other areas of social policy, would exist only at a national level. National governments would be responsible for decision making and obtaining the requisite information on outcomes. In that scenario, change, including greater influence for poor people, would derive from changes in political processes and participation at the national level. In recent years a major concern of aid donors has been to support as much as possible the processes of domestic politics and domestic accountability. Aid instruments such as 'direct budget support' have been developed to enable donors to feed into domestic budgets rather than to enforce their own priorities.

However, despite donors' concern with national level processes and politics, they retain much influence over how aid is spent. Donor governments, foundations, NGOs and multilateral agencies still make choices about where to give aid – to what countries, to what sectors, and through what specific projects and instruments. These decisions are not subject to the usual feedback mechanisms of politics (Barder, 2009). Rather, the feedback mechanism, where there is one, is between the decision makers in donor countries and their domestic taxpayers who want to be reassured that their taxes are being well spent.

There is thus something of a conflict, and a balance to be struck, between democracy in donor countries, where citizens rightly demand accountability for how their money is spent, and

democracy in recipient countries, where people need to be able to influence how money is spent in order to ensure that it really does deliver the outcomes they want (Box 1).

Box 1 – Aid, national government spending and the priorities of the poor

Much evidence suggests aid, national spending notwithstanding, is misaligned with what the poor want and value. First, it is clear that aid often does not go where, on the face of it, need is greatest. A clear example is agriculture. As highlighted in the 2008 WDR, *Agriculture for Development,* three of every four poor people in the world continue to live in rural areas where most depend on agriculture for their livelihoods. At the same time, aid for agriculture declined over the 1990s, until the early 2000s.[1] While presently many countries remain highly dependent on Overseas Development Assistance (ODA) to agriculture, donor interventions are described as "fragmented, overlapping, discontinuous, and sometimes contradictory" (p. 257). Moreover, as a recent Gallup World Poll conducted in sub-Saharan Africa made clear, aid is not spent on the MDGs that people prioritize most (Tortora 2006) – for example far more is allocated to primary education rather than jobs. This discrepancy is all the more acute in the case of highly aid dependent countries, where aid represents two-thirds of financial flows; because it is largely filtered through the MDGs, it flows to issues associated with them – excluding a priori spending on infrastructure, for example (and till 2005, employment).

But these instances of a likely mismatch seem to be symptomatic of a bigger problem: the absence of data in nationally representative household surveys on many aspects of development that people value, according to participatory accounts of poverty such as *Voices of the Poor.* Alkire (2007) compares the dimensions that emerge as important in *VoP* with those collected in well-known large-scale survey instruments designed to measure deprivation namely the World Bank Living Standards and Measurement Survey (LSMS), the World Bank Core Welfare Indicators Questionnaire (CWIQ), the US Agency for International Development Demographic and Health Survey (DHS) and the UNICEF Multiple Indicator Cluster Survey (MICS). She concludes that empowerment, physical safety, psychological wellbeing, employment quality and the 'ability to go about without shame' are largely absent.[2] It would therefore be extremely difficult to evaluate the impact of aid, alongside other interventions, on these dimensions.

The fear among many observers of development is that donors' focus on results will tilt this balance more towards taxpayers in donor countries, and potentially away from what delivers the best outcomes for poor people as they see it, and from what strengthens domestic accountability in developing countries – since developing country governments also are concerned with donor approval.

However, neither of these are a necessary consequence of the drive for more information on results and outcomes. If outcomes are defined according to what poor people themselves want, then the same results framework could deliver accountability to both taxpayers and poor people. But defining the outcomes properly is crucial. In the absence of an easily communicated way of articulating what poor people most value, the default position will be to reach for outcome

[1] http://www.oecd.org/dataoecd/54/38/44116307.pdf.

[2] OPHI's Missing Dimensions research programme is aiming to devise internationally-comparable survey modules that could be integrated in traditional household surveys to fill these gaps.

measures like 'numbers of schools built'. Many years of experience have shown that such 'intermediate' measures may not provide the right incentives to donors, recipient governments and communities to create lasting change, because they may not necessarily reflect what poor people would themselves define as the most important outcomes.

An improved way of defining and measuring development outcomes could potentially shorten the distance between donors and poor people. In doing so it would soften the contradiction donor governments face of having to be accountable to both tax payers and aid recipients, and provide a better metric for understanding if what aid and development are doing in the name of improving people's lives is actually having any effect. It could assist both donors and the national governments of developing countries as they try to allocate resources in ways that achieve the best outcomes. But, in order to be effective, it has to be a metric that is useful and useable for policy makers. Reconciling the two is not a simple matter.

3 A brief history of development outcome measures

It is possible to read the story of development outcomes as an ongoing effort to marry researchers' evidence on complexity and interconnectedness, poor people's own complicated experiences of poverty, and politicians' needs for simple and easily understood measures of progress.

The treatment of economic indicators of progress exemplifies this difficulty. The most commonly used outcome measure for progress in all countries is monetary – whether in the form of measuring income or consumption per head from survey data or of calculating Gross Domestic Product from national accounts. However, as research has become more sophisticated, it is increasingly clear that, when asked, poor people are likely to define economic poverty as much in terms of asset poverty as income poverty. In *Voices of the Poor,* for example, Narayan *et al*. (2000) observe: "The poor rarely speak about income, but they do speak extensively about assets that are important to them" (p. 39). Researchers have responded to this gap, and thanks to huge investments in survey methodologies and data collection, information about the assets held by poor people, and how this is connected to their experience of poverty, is now widely available. But despite this being closer to poor people's own understanding of poverty, most economic analyses of poverty still focus largely on income. In part this is because it is much easier to express incomes in a single comparable unit – the US dollar, expressed in terms of its purchasing power parity. Calculating and comparing asset holdings can be more difficult, though not impossible – e.g. (Booysen, Berg, Burger, Malitz, & Rand, 2005), Filmer and Pritchett 1998, Sahn and Stifel 2000.

However, the biggest shift in development outcome measures in the last thirty years has been a move to include non-economic as well as economic measures of progress. Dudley Seers, the first director of the Institute of Development Studies at the University of Sussex, argued more than forty years ago for the "dethroning of GDP" – that rather than focusing on economic growth, a country's development performance should be assessed against other measures such as infant mortality, inequality or unemployment (Seers, 1967). More recently, Amartya Sen has pioneered the shift towards an understanding of 'development as freedom' (Sen, 1999) or the enhancement of people's

capacity to advance valued goals – an understanding that underpins the UNDP Human Development Reports. Much recent work on development outcomes is explicitly informed by Sen's approach.

Non-economic indicators, such as infant mortality rates or primary school enrolment ratios, are now widely used alongside income to measure development outcomes. What has come to be called the 'human development approach' has informed a great deal of development theory and practice since it was given a name and an institutional home in the first of the UNDP's Human Development Reports (HDR) (UNDP, 1990). The key innovation in the 1990 HDR was the construction of the 'human development index' (HDI), in which indicators of living standards, health and education were aggregated into a single number. This simplicity is the reason for its power and influence, but also a source of criticism from some researchers, who argue that the aggregation masks a great deal of complexity and involves many assumptions about how the different components of the index should be weighted against one another (Ravallion, 2010).

The Human Development approach also informs what are probably the best known development outcome measures: the Millennium Development Goals (MDGs). Described as 'human development meets results-based management' (Hulme, 2010), the goals were developed through a process of international summitry and negotiations led by OECD donors and the United Nations. The targets attached to the MDGs include halving the percentage of people who live in extreme income poverty, reducing the infant mortality rate by two-thirds and getting all eligible children enrolled in school by 2015.[3] The MDGs continue to define development progress for many donor governments and agencies, and some developing country governments (Fukuda-Parr, 2010).

The goals are a reasonable set of objectives, have commanded wide political engagement, and are probably a factor behind the growing public support for aid in many donor countries between 2000 and 2005, and the consequent rise in aid budgets. However the lack of input from poor people in their creation is evident too. Many issues that would be of key importance to poor people – such as the need for infrastructure and threat of violence – are not mentioned in the MDGs.[4] And the goals that receive the most attention and resources are not necessarily those that poor people prioritise. When asked, as they were in the Gallup World Poll (Tortora, 2009), people in sub-Saharan Africa ranked providing jobs for young people as the fourth most important of the MDGs, ahead of universal primary education and other goals that typically receive more attention from donors.

The difficulty of incorporating the needs and priorities of poor people into goal setting and measurement in development has not escaped the attentions of researchers. Starting with the pioneering work of Robert Chambers at the University of Sussex (Chambers 1983, 1993, 1997, 2007), different methods have been investigated, from different empirical and theoretical perspectives, to understand, report and use better the views and preferences of poor people in making and evaluating development policy.

[3] For the complete list of indicators, see: http://mdgs.un.org/unsd/mdg/Host.aspx?Content=indicators/officiallist.htm.

[4] The World Bank's 'Voices of the Poor' study, which involved 60,000 people in 60 countries found that poor people tended to be more focused on assets than incomes, and that employment, transport and water came up most frequently as high priority problems (Narayan, 2000). With the exception of water, none of these featured in the original MDG and employment only became a target in 2005.

Two important approaches to poverty measurement which attempt to incorporate poor people's own views are wellbeing approaches and participatory poverty assessments.

- Wellbeing approaches use surveys and other research methods to identify the component parts or 'dimensions' of wellbeing. These are generally said to comprise physical, psychological, relational and spiritual components appropriate to the particular context, which can be used to measure people's satisfaction with different aspects of their lives (Gough & McGregor, 2007). Individual wellbeing can be expressed quantitatively, and the different components can be assessed separately or added together to give a single indicator.
- Participatory poverty assessments are based largely on qualitative research and use a variety of methods to assess how people in a given locality define poverty and their priorities for change (Norton & al, 2001).

These approaches have to some extent traded off usability for comprehensiveness. They typically involve complex questionnaires in addition to qualitative research, and produce results which in some cases are not directly comparable to results from other communities or countries because they incorporate context-specific measures of what matters to people.

Both approaches have at different times been adopted by governments as useful inputs into policy making. PPAs are the longest established and have been widely used by both donors and governments as one-off exercises for programme planning purposes. Robert Chambers estimates that PPAs have taken place on every continent (Chambers, 2007), and they have generated some important insights such as the importance of time poverty among women, or of seasonal patterns of poverty among agricultural communities.

The wellbeing approach is catching on in domestic policy circles in donor countries as well as in development, and again has generated some important insights such as the importance of religious belief to many people's wellbeing. However, a leading researcher on wellbeing has recently argued that the approach is not necessarily appropriate for answering the type of questions about individual policies or programmes that an aid official, NGO or civil servant might want to know. Instead, it provides a more general framework for evaluating the circumstances and progress of individuals or communities (White, 2009).

Despite some limited usage by governments, the failure to translate the insights gained from these approaches into changed donor practice is also widespread. An example is the fate of 'Voices of the Poor', almost certainly the largest PPA ever carried out. The study was funded by the World Bank, which should have made it influential among other donors. It repeatedly mentions improved roads and better jobs as two outcomes of high priority for many of people interviewed. Yet it was not until many years after this research that donor spending on infrastructure began to rise, and jobs are only just coming on to the mainstream international development agenda, with recent reports from the UN and a World Bank World Development Report on employment scheduled for 2013.

The challenge for those trying to balance policy influence, adequate representation of poor people's values and methodological rigour is to build on existing approaches to measuring poverty in a way which reflects the reality of poor people's lives to develop new outcome measures which can be more easily used for answering the policy questions demanded by both donors and national governments.

For policy makers, the key to the usefulness – or not – of different methods is simplicity. The ideal is a measure which can be expressed as a single number, but which contains a great deal of information within it, is comparable across different countries, regions or people, and can be regularly monitored over time. Income per head, expressed in a single currency such as the US dollar – and corresponding income poverty measures – is hard to better from the perspective of simplicity, one reason for its longevity as an indicator of development despite many limitations. The Alkire Foster measure, and its application in the UNDP's Multidimensional Poverty Index (MPI), represent an important advance in this respect (see Box 2).

The current interest in results has been extremely useful in focusing attention on how development outcomes are defined and measured. A great deal of work has already been done in this area, providing a good basis on which to develop an indicator that both meets the needs of this generation of policy makers to define and measure results, and addresses the concerns of earlier researchers about providing an understanding of development outcomes drawn from the views, experiences and priorities of poor people.

Box 2 – The Alkire Foster Poverty Measure and MPI

There is something of an inverse relationship between the extent to which measurement tools capture the range of issues that are important to poor people and their ease of use for policy makers. The wellbeing and PPA approaches fall on the side of trying to capture complex realities, while the Multidimensional Poverty Index (MPI), by contrast, aims at simplicity. By aiming to compare poverty across developing countries, the MPI necessarily draws on a thin comparable body of data – the index is highly circumscribed in the range of indicators that are available, thereby excluding some issues known to be important to poor people. But at the same time, it provides policy makers with a single figure for each country, community or household that is sensitive to both the average extent and depth of deprivations among the poor (Alkire and Santos 2010). The commitment by the designers of the index to using widely available and comparable data means that a single number can be calculated for the percentage of people living in poverty as defined by the MPI in a given country or region, which can then be compared with other regions or countries. This is where the choices made which exclude issues of interest to poor people pay off in terms of usability of the measure. The Alkire Foster methodology, however (on which the MPI is based) provides a means of incorporating a broader range of indicators and weights into the measurement of multidimensional poverty which can be country specific. Several governments have adapted this tool to their own particular context – notably in Bhutan and Colombia.[5]

4 Measuring outcomes in health care

The quote at the beginning of this paper illustrated that development is not alone in grappling with the question of how to know if the outcomes achieved are those that will do the most to improve people's lives as they see it. Other areas - for example, health care - have also sought to answer this question. Methods and measures have been developed which have gained much more traction with

[5] http://www.ophi.org.uk/policy/national-policy/.

policy makers than those used in the development sector, and which are being widely applied. Looking at approaches developed and used in the health sector might provide useful insights into how to move beyond existing approaches in development, and in particular, how to reconcile the twin demands of complexity and simplicity.

It is perhaps unnecessary to note that health care differs from development. However, there are general parallels, particularly in terms of the need to evaluate and measure performance and impact, the need (in taxpayer-funded systems like the English National Health Service (NHS)) to account for money spent, the desire to involve and take account of users' views and values, and the general objective to be cost effective. This last may seem a rather narrow economist's view, but it captures a key reason for a need to develop measures of outcome (the 'effect' part of effectiveness). How effect or impact is and should be measured is, as we will see, important at all levels of decision making in health care and, with careful design, can reflect what health care and development are all about - improving the conditions of people's lives. The full quote from Don Berwick cited at the beginning of this paper is as follows:

> *"The ultimate measure by which to judge the quality of a medical effort is whether it helps patients (and their families) as they see it. Anything done in health care that does not help a patient or family is, by definition, waste, whether or not the professions and their associations traditionally hallow it"* (Berwick, 1997).

The sentiment underlying this quote, with appropriate substitutions ('development' for 'health care' etc), is equally hard to dispute for development.

So, what has health care been doing to develop, in Berwick's words, the 'ultimate measure'? And how might this experience be useful in trying to develop similar measures for development?

Drawing primarily on the experience of health care in the UK (more particularly, the English NHS) as well as the wider issue of clinical evaluation, below we describe the development of patient-based health care outcome measures (PROMS) and their use as performance measures. We then discuss how these data inform decisions regarding resource allocation at all levels in health care, with a focus on the way that the National Institute for Health and Clinical Excellence (NICE) use them to make recommendations about health care technologies. First, however, some contextual history.

4.1 Relieved, unrelieved and dead

It is nearly a century and a half since Florence Nightingale first devised a simple, tri-dimensional outcome measure for her patients – relieved, unrelieved and dead (Nightingale, 1863). Nightingale's outcome measure may not seem particularly revolutionary. Indeed, it may strike many as a somewhat crude summing up of the impact of health care on a patient's health status. However, for most of its history, the National Health Service in Britain (not uniquely among the world's health systems) produced an even cruder measure of outcome for hospital care: discharges and deaths combined. It is only a slight exaggeration to say that for most of its life, the NHS, at least in a statistical sense, did not know whether its patients were discharged dead or alive.

This should not suggest a history of deliberate wilful disregard for patients or the outcomes of medical interventions. Clinicians have, of course, developed measures to guide and inform their clinical practice, which provide important and relevant information about the impact of health care

interventions on clinically-defined variables. But while these are useful, they typically fail to inform wider questions crucial to measuring the *overall* impact or effect of health care on patients' quality of life.

Moreover, the multitude of clinical indicators used in medical practice do not always distinguish the aspects of health patients consider important, or their relative value to patients. The surgeon's observation that the operation was a success, but the patient died may be an example of the dark humour of the medical profession, but it is nonetheless indicative of a gap that can exist between clinical and patient views of what matters in health care.

In addition, multiple clinical measures do not facilitate comparisons of health impact across clinical areas, so do not reveal anything about the overall effect of spending and service delivery across different disease areas and their ultimate effects on health outcomes. And they do not help identify how those outcomes may be improved by different allocations of resources between services and patients.

Over the last few decades, the NHS has come a long way in the number, type and sophistication of performance measures it collects and collates in order to inform decisions from clinical choices for individual patients to high level resource allocation. The counting of process measures such as the number of patients treated, operations performed, prescriptions written and so on still forms the core of its performance monitoring, but over the last few years (and uniquely for a national health system) there has been a revolution in the way it measures its impact on patients' health and health-related quality of life. Patient reported outcome measures (PROMs) are now beginning to capture important aspects of health care's impact on the things that patients most value - their mental wellbeing, their ability to carry out normal physical activities, the degree of pain and discomfort they endure and so on.

The final report of the NHS Next Stage Review observes, 'Just as important [as patients' experience of health care] is the effectiveness of care from the patient's own perspective which will be measured through patient-reported outcomes measures (PROMs[6]).' (Department of Health, 2008).

The development and use of PROMs in the health service in the UK has expanded rapidly in recent years. Two main types of PROMS are in use – generic measures which aim to capture outcomes of relevance to any health related intervention, and specific measures detailing outcomes of specific treatments such as for knees or hips. Generic measures are used by the National Institute of Health and Clinical Excellence (NICE) as an input into resource allocation decisions, while the NHS is employing a combination of generic and specific measures to assess and compare the results of clinical interventions by individuals or institutions.

4.2 Rationing in a NICE way

While various types of PROMs have been used for many years as part of the battery of outcome measures used in clinical trials - alongside biometric measures, for example - and as part of general regulatory requirements to ensure safety and effectiveness – the creation of NICE in 1999

[6] It is worth noting that the term PROM is, as Browne et al (2007) point out, something of a misnomer. PROM instruments do not directly measure the outcome of a health care intervention but rather the health related quality of life of an individual at a specific point in time. Changes in patient-reported health (e.g., before and after surgery) allows us to work out the effect or outcomes of surgery, after controlling for other factors.

introduced more formalised requirements to collect PROMs data as part of its methodology for carrying out economic evaluations of new and existing therapies. In broad terms, NICE's role was and is to make recommendations to the NHS about which health technologies are good value for money. The use of PROMs data for making resource allocation decisions took the use of patients' perspectives into new territory for the NHS. It is an experiment which has been watched with interest by health services around the world, several of whom are setting up similar bodies.

Deciding which health care options represent best value for money depends on being able to weigh up the benefits and costs of each. Given a fixed budget, spending on one option means those same resources cannot be used in another way. That is, every decision carries an opportunity cost – the benefits that would have been possible from the next best alternative use of them. Weighing up the benefits possible from each potential option imposes the important requirement that the benefits be measured in a commensurate way. The benefits from quite dissimilar options – for example, hip replacement versus mammography screening – need to be measured using the same 'metric'.

This task requires an evaluation of the costs and effects (outcomes) of different therapies. The definition of the 'effect' adopted by NICE is an outcome measure which combines the two key objectives of any health care intervention - an increase in the length and quality of life. The measure, known as the quality adjusted life year (QALY), weights the outcome (years gained) of an intervention by the health-related quality of each extra year of life.

QALYs are estimated by applying a quality weighting to the length of life lived in a given state of health. The quality weighting in turn is derived from a questionnaire known as the EQ-5D. The questionnaire solicits and aggregates patients' self-reported health along key dimensions of health related quality of life, which are then weighted relative to one another on the basis of data collected from the general public.

Important to note here is that the quality weightings (or values, or 'utilities') range from 1 (full health) to 0 (a quality of life so poor it is as bad as being dead). Health states that are worse than being dead are possible, and are represented as weights < 0.[7] Because the weightings are derived from survey data, they reflect the values of the general publicrather than those of health professionals or patients themselves. More details on the methods used to derive these weights are provided below.

For example, if someone lives for 10 years in full health, 10 x 1 = 10 QALYs. However if the quality of life is less than perfect – for example, 0.5, then each year of life is 'worth' only half a year of full health. Ten years lived in a quality of life of 0.5 will be equal to 5 QALYs.

Assessing value for money in health care involves evaluating the change in QALYs that is caused by treatment relative to the cost of achieving those QALY gains.[8] For example, on current standard treatment, a person may have expected to experience:

[7] In fact in the UK value set used by NICE, nearly one third of the EQ-5D states have a value < 0, i.e. have a negative value/weight. That means a year lived in such a state actually reduces your overall QALYs. And a treatment that takes you from a negatively valued state to a positively valued one can give you 'super QALY gains' such that each additional year of life will be worth more than one QALY.

[8] For any given state of quality of life, we assume that the utility in any one period is the same as in any other, and that that is independent of the order in which health states are encountered (whether as part of a stable,

- 20 years life expectancy x 0.5 quality of life = 10 QALYs

And with a new treatment, quality of life might increase:

- 20 years life expectancy x 0.9 quality of life = 18 QALYs

The change in QALYs is (18-10) = 8. If there is a net increase in costs to the NHS of providing this new treatment of £20,000, we can estimate the change in costs divided by the change in QALYs:

- £20,000/8 QALYs = an incremental cost per QALY gained of £2,500

This incremental cost effectiveness ratio (ICER), expressed in terms of the change in cost per QALY gained, can be compared across very different uses of health care budgets. Options with low ICERs represent better value for money than options with high ICERs. These approaches are used routinely by NICE, using a carefully specified set of methods (NICE 2008a), as a means of assessing the value for money of new health care technologies. NICE uses a 'rule of thumb' to judge whether or not any given ICER represents good value for money: it has a 'cost effectiveness threshold' of GBP 20,000 – 30,000 per QALY gained. Technologies with ICERs below this level are more likely to be recommended; those with ICERs above that range are more likely to be rejected. Criteria other than cost effectiveness are taken into account as well (NICE 2008b; Rawlins et al 2010; Devlin et al 2011.[9]

The use of QALYs (rather than say, lives saved, or life years gained) by NICE facilitates a universal comparison between treatments - theoretically between different treatments for different conditions, although this is not the sort of comparison NICE does, at least, not directly. In other words, a 'QALY is a QALY is a QALY' no matter which health care intervention produces it. In theory at least, NICE's recommendations should help the NHS make decisions about how best to spend its limited and finite budget each year in order to move towards improving the value each health care pound achieves - at least in terms of QALYs.

Perhaps unsurprisingly, NICE has generated considerable controversy as a result of its its recommendations to the NHS as to the 'best buy' interventions. . This is in part due to the very difficult nature (and impact) of these decisions, which in some cases effectively deny treatment of (some) likely positive benefit to some patients because the health technology was not cost effective enough and hence deemed to be unacceptable value for money forthe NHS. Moreover, given the impact NICE decisions can have on pharmaceutical sales, not just in the UK, but in many other countries which can also base purchasing decisions on NICE evaluations, it is also unsurprising that its methodology has come under intense scrutiny. This scrutiny has focused on, among other issues, the chosen outcome metric of the QALY and the use of the EQ-5D measure as the quality weighting.

For example, from time to time concerns have been expressed about whether the EQ-5D captures all aspects of patients' quality of life that are relevant in some conditions. While there are challenges in

---

improving or deteriorating sequence of health states) and how long the state is experienced for (a week, a year, 10 years...). These are the assumptions of additive separability and constant proportionality.

[9] For instance, according to Rawlins et al (2010), they take into account: severity of the illness, end of life, stakeholder persuasion (e.g., about quality of life effects that are missed in the QALY calculations), significant innovation, disadvantaged populations and children.

measuring health related quality of life in, for example, patients with hearing and vision problems, in most cases the EQ-5D does seem to be a broadly reasonable measure of health (Wailloo et al 2010). Also, there has been debate about whose values should be used to determine the quality of life weights, and whether this should be the general public (as is current practice) or those of patients themselves (Brazier et al 2005, Brazier et al 2009).

While NICE's role in evaluating new and existing health technologies in terms of their cost effectiveness continues, a development in the English NHS begun in 2009 expands the use of patient reported outcome measures.

4.3 PROMs and the NHS: Routine measurement of patients' quality of life

Since April 2009, patients undergoing one of four surgical procedures (hip and knee replacements, varicose vein removal and hernia repair) are asked to complete a set of questionnaires before and after surgery designed to elicit information about their health related quality of life (DH 2008).

The objective of this survey exercise is to go a step further in enabling the NHS to understand better the link between the way services are delivered and patient outcomes; to monitor the performance of specific individuals and institutions in providing similar types of care; and to improve the information available to patients when making decisions about their treatment.

Box 3 below briefly summarises the English NHS PROMs initiative.

Box 3: Patient reported outcome measures in the English NHS

The PROMs initiative collects information from patients (before and after their operations) using an appropriate disease-specific questionnaire (for example, the Oxford hip and knee score[i]) and a generic questionnaire given to patients in all four procedures. Completion of these instruments is voluntary.

The generic instrument is a particularly important aspect of the NHS PROMs initiative. While a number of generic health questionnaires[10] have been developed over the years, the instrument of choice in the NHS, as for NICE, is the EQ-5D.[11]

Data from the PROMs initiative is published routinely by the Department of Health - it details average disease-specific and generic PROMs scores both before and after operations at the level of individual hospitals (both NHS and private, for patients treated privately but paid for by the NHS). Over the two years beginning in April 2009, of around 484,000 eligible patients, 329,000 returned the pre-operative and 215,600 the post operative questionnaires. The average scores for each hospital are adjusted to take account of differences in the case mix treated.

Figure 1 presents actual data from the PROMs initiative, showing for English NHS hospitals average changes in the EQ-5D index following a hip replacement operation. All hospitals produce gains in the

[10] For example, the Short Form 36 (SF-36) (and its reduced version, the SF-12, and the SF-6D which is accompanied by utilities; The Health Utilities Index (HUI): the Assessment of Quality of Life (AQol); and the 15D – for a review, see Morris et al (2007). [

[11] The EQ-5D is a trade mark of the EuroQol Group (http://www.euroqol.org/eq-5d/what-is-eq-5d.html) . While the PROMs programme and NICE currently use the EQ-5D, there is interest in both cases in adopting the expanded level version of it, the EQ-5D-5L, which is now available.

index, some more than others (although none are statistically different from the national average gain).

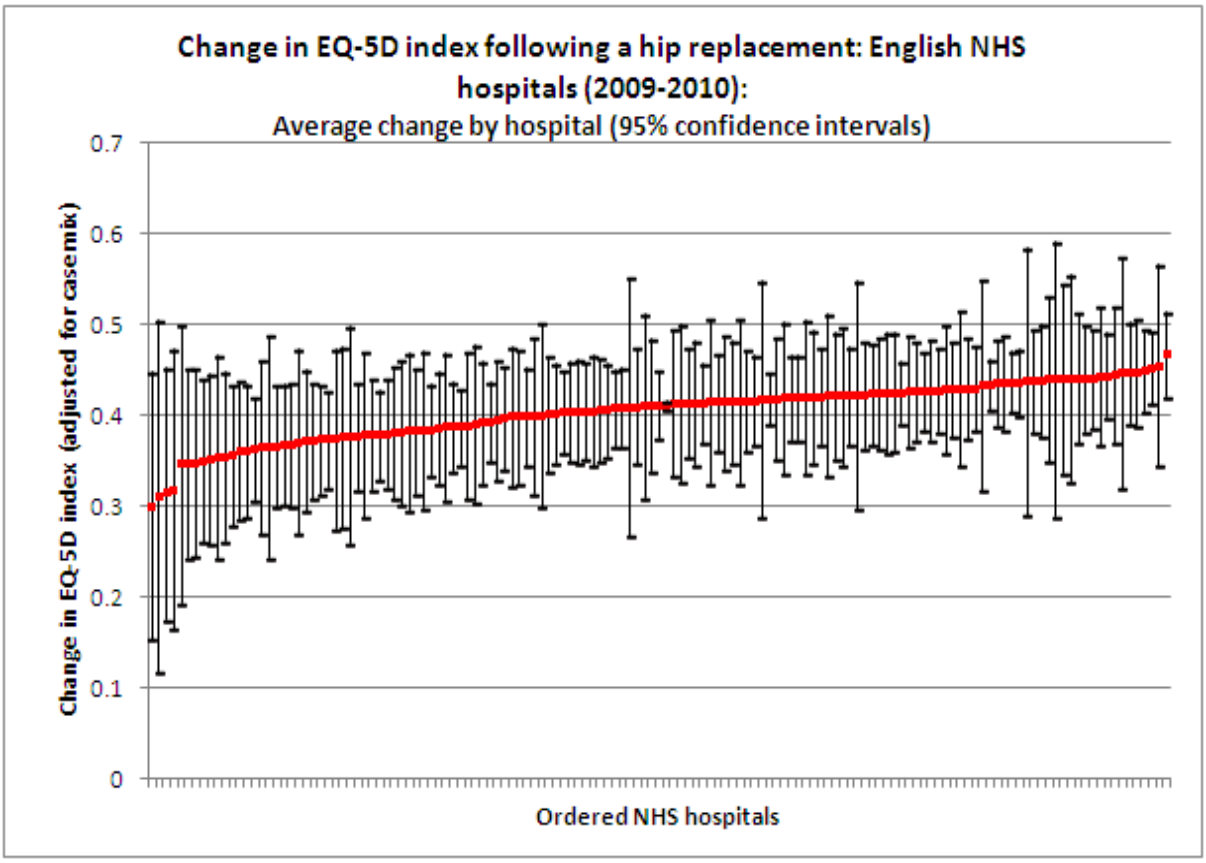


Figure 1 – Change in EQ-4D index following a hip replacement based on NHS data

Although data from the PROMs initiative are publicly available (HESonline 2011), they currently remain classified as 'experimental statistics' by the Department of Health. While the hospital comparison section of the NHS Choices website provides comparative 'health gain' EQ-5D index data, and most of the focus of analysis and use of these data to date has been on identifying and understanding differences between providers' performance. . Nevertheless, the anticipation is that alongside other performance measures and routine data collected as part of patients' records, PROMs will enable patient perspectives to be taken into account in key aspects of the NHS, including:

- Informing patients' choices of treatments and providers
- Linking provider payment to their performance in improving patient health
- Understanding and managing referral from primary to secondary care
- Facilitating a dialogue between clinicians and managers about the delivery of care
- Use by health care professionals to monitor and improve health care practices
- Regulating for safety and quality in health care services

4.4 Debates on PROMS

While there has been little or no arguments about the potential (and actual) benefits of measuring patient assessed health outcomes, as Brazier and Longworth have noted (2011), the *development* and *use* of PROMs in health care has not been without debate or argument about, for example, the best way to measure patient assessed outcomes, whether the focus should be on health-related outcomes or broader measures of well being, how to turn descriptive health states into single overall indices and which methods to use to value health states and whose values to use. These issues reflect general questions about the validity, reliability and responsiveness of health status questionnaires (see Box 4).

Box 4 – Validity, Reliability and Responsiveness

Validity
Content validity: Does the instrument exclude dimensions of health that are important to patients?
Face validity: Are instrument items relevant and appropriate for the population?
Construct validity: Does the resulting measure reflect known differences between groups?

Reliability
Can the measure reproduce the same value on two separate administrations when there has been no change in health? This can be over time, between methods of administration or between raters.

Responsiveness
Does the measure capture "clinically significant changes" in health?

On responsiveness, for example, it has been argued that the EQ-5D fails to capture relevant health gains following cataract replacement operations in part due to ceiling effects (cf Ferreira et al, 2008). In a pilot of the EQ-5D prior to the introduction of this measure in the NHS, a significant number of patients undergoing cataract surgery recorded pre-operative EQ-5D scores equal to 1, or perfect health, raising the question of whether it is an adequate measure of health related quality of life in the case of patients with cataracts (Devlin et al 2010)

Given the original aims of the developers of the EQ-5D, compromises between the complexity and length of questionnaires and ease of use have had to be made. A key question has been whether the deficiencies of generic instruments such as the EQ-5D are worth the benefits of recording such data and in particular, given the use to which the data might or will be put, the extent to which they matter.

5 Can these methodologies be translated to development?

The need for greater public inputs into decisions about both resource allocation and measures of performance and improvement is just as great in development as in health. But unlike in development, serious and systematic attempts are underway in health to collect and use this information. Could this approach be useful in defining and measuring development outcomes so that they are better informed by the views and priorities of poor people? Below we describe in more detail how these measures are constructed in the health sector, and some possible issues involved in adapting them to measure development outcomes.

The construction of outcome measures in the health service has involved answering two key questions:

- How do we *measure* the overall quality of life and changes in that, experienced by patients?
- How do we *value* the overall quality of life, and changes in that, experienced by patients?

The analogue, in development, is:

- How do we *measure* the overall quality of life, and improvements in that, experienced by local people as a result of investments in development?
- How do we *value* the overall quality of life, and changes in that, experienced by local people as a results of investments in development?

The remainder of this section describes how these questions were answered in the construction of outcomes measures in the health service, and considers the implications for development.

5.1 Defining and measuring outcomes

If the goal is to measure the improvements in outcomes produced by spending on development, regardless of whether that spending is directed toward health care, agriculture, infrastructure or other initiatives, this suggests a requirement for a 'generic' measure. By identifying the principal outcomes which are seen as important to those for whom development is meant to benefit, a generic outcome measure can provide a way of capturing and comparing the outcomes possible from very different uses of resources. The dimensions would form the basis for a questionnaire, which could be used with local populations which are the focus of development efforts, to measure and evaluate the outcomes of aid-funded interventions from their point of view.

The history of generic outcomes measures in the health sector

In the health care sector, Rosser and Kind (1978) undertook pioneering work to develop a generic outcomes measure. Any given health state was described in terms of just two dimensions: *disability* and *distress.* The degree of disability was measured in 8 levels, from 'no disability' (level 1) to 'unconscious' (level 8). Distress comprised four levels: 'no distress' through to 'severe distress'. Each of the 'states' described by the combinations of these dimensions and levels were assigned a value (quality of life weight) on the 0-1 scale described earlier, as required in the estimation of QALYs.

As interest in outcomes measurement continued to grow, more instruments were developed. Some of these focussed exclusively on measuring health outcomes, rather than providing an input into economic evaluation. For example, the SF-36 (Brook et al 1979) is a widely used generic measure of health outcomes, with 36 questions on various aspects of physical and mental health. A large number of disease specific instruments were also developed – again, generally with a focus on outcome measurement rather than facilitating assessments of value for money.

However, the increased demand for evidence on cost effectiveness in health care also led to the development of measures suited not only to measuring outcomes associated with particular interventions but also to giving relative weights (or 'utilities') to the different aspects of quality of life improvements which they produce. Examples include the Health Utilities Index (Horsman et al 2003), the Assessment of Quality of Life (Hawthorne et al 1999), and the EuroQol Group's EQ-5D

(Brooks 1996). The EQ-5D, as discussed above, is the generic instrument NICE recommends for use in evidence submitted to it (NICE 2008). It is included in the NHS PROMs programme (Devlin and Appleby 2010) and widely used internationally.

How are outcomes measures developed?

Developing a questionnaire ('instrument') to measure the perceived quality of outcomes comprises four key stages.

First, the underlying 'measurement model' needs to be established. That is, the instrument to be developed should be based on an underlying theory, so that there is clarity about the concepts to be measured and how these relate to the purpose of measurement. For example, relevant theories in health might relate to utility, quality of life, needs, subjective wellbeing, 'capabilities' and so on.

Second, research is undertaken to generate the instrument's content i.e. questions ('items'). Streiner and Norman (2008) suggest a first step of checking the literature for evidence on any existing instruments and the items they use, while also cautioning that terminology in previous instruments may be outdated, and that previous instruments may be inadequate for a variety of reasons. Instrument development will therefore involve a careful process of identifying potential items, from sources that might include patients/research subjects, clinical observation, theory, research and expert opinion. McKenna (2011) emphasises that the content of quality of life questionnaires should always come from patients, specifically suggesting that both the items and their wording should be based on qualitative analysis of interview transcripts with (at least) 30-35 patients. In contrast, Streiner and Norman (2008) note there is no set number of people who should be interviewed. Instead they describe a process of either structured or unstructured interviews to identify relevant themes, depending on how much prior evidence is available, with a common approach being to 'sample to redundancy' i.e. to interview people until no new themes emerge. Streiner and Norman (2008) also refer to a wider set of approaches to obtaining patient input, such as focus groups, both to establish relevant general themes to be used by researchers in developing items and subsequently to check that items are relevant, clearly worded, unambiguous, readily understood and that all the main themes have been covered.

The identification of potential items will generally suggest a large number of possible questions for inclusion in the instrument. The third stage therefore entails a process of content refinement and item reduction. Parsimony is desirable in order to avoid an unnecessary burden on those asked to complete questionnaires. Where the instrument is to be accompanied by a value set (weights, or 'single index' numbers), this imposes a rather stricter requirement to minimise the number of items that would be involved in valuation exercises. The process of item reduction entails making sure items are relevant; clearly expressed; avoid duplication; and are capable of registering change over time. Understanding how items relate to each other can be formally explored through such methods as principle component analysis, multiple correspondence analysis or fuzzy set theory.

Finally, the instrument should be subjected to scaling and psychometric testing. This entails testing the draft questionnaire with a set of relevant patients to check their ability to understand it, complete it and that the items are considered relevant, and formal psychometric tests of reliability and validity (which we describe below).

The development of the EQ-5D, described by Williams (1990) and Kind, Brooks and Rabin (2005), provides an example of this process. The underlying conceptual basis was the measurement of health related quality of life, defined as 'the value assigned to duration of life as modified by the impairments, functional states, perceptions and social opportunities that are influenced by disease, injury, treatment or policy' (Patrick and Erickson 1993). The development process was guided by the objective of producing a short, readily self completed measure of a common core of dimensions of health-related quality of life capable of yielding a single index value for any given health state it defined (so as to facilitate its use in estimating QALYs). An initial six EQ-5D items were selected following a review of existing health status measures (Quality of Wellbeing Scale, The Sickness Impact Profile, Nottingham Health Profile and the Rosser Index): mobility, self care, an ability to perform one's main activity and also family and leisure activities, pain and discomfort, and anxiety/depression. These were later tested in a 'lay concepts of health' survey (van Dalen, Williams and Gudex 1994) with members of the UK general public. This suggested that energy/tiredness should be added to the six dimensions – and to accommodate this, the two dimensions of 'main activities' and 'family and leisure activities' were combined ('usual activities') as the latter had been observed to add little to the overall valuation of states. Similarly, energy/tiredness also appeared in later testing to contribute little to health state valuations and was dropped. This led to the version of the EQ-5D which is currently in use.

The EQ-5D consists of five questions that are generally applicable regardless of, for example, the type of illness a patient is suffering from or the particular medical intervention he or she receives. Extensive testing and validation in many patient populations around the world has resulted in a fairly simple set of questions covering the five included dimensions. Each of these five 'dimensions' can be responded to in three ways - no problem, moderate problems and extreme problems. The five dimensions and three possible responses give rise to 243 different possible health states (3 levels to the power of 5 dimensions). In addition, using a thermometer-type scale (the Visual Analogue Scale - VAS) running from zero (dead) to 100 (perfectly healthy),patients are asked to make a mark representing their view of their overall health. (see Appendix 1 for a sample of the EQ-5D questionnaire).

How do we judge the quality of an outcomes measure?

Good outcomes measures must meet many requirements. For example, the quality of a health outcome measure may be judged in terms of psychometric criteria. These include whether the measure is (a) *reliable* – does it produce consistent measures? And (b) is it *valid* - does it measure what it is meant to measure? Validity might be further defined in terms of (a) *criterion validity* – the extent to which it correlates with other measures that are known to be valid; (b) *construct validity* – how the instrument's results relates to others suggested by theory; (c) *content* or *face validity* – whether the instrument appears to contain and cover the concept it is measuring. (Streiner and Norman, 2008). Other criteria to be considered include whether the instrument is *responsive* – that is, the extent to which the measure of outcome is sensitive to changes in health. And *feasibility* concerns whether or not it is possible to use the instrument in practice.

There are additional criteria for measures to be used in economic evaluation (Morris et al 2007). First, they should provide an unambiguous measure of benefit. It should be possible to summarize any given health outcome described in terms of the various dimensions and levels by a single number. Second, the measure should be capable of comparing different possible allocations of

scarce resources. This emphasises, as noted above, the requirement for the outcome measure to be 'generic' in nature. Third, it should be capable of being interpreted in terms of value, since this economic concept lies at the heart of the assessment of value for money. The implication of these three criteria for the development of an outcomes measure is that it should be generic in nature and capable of being summarised by a number that reflects people's values and preferences.

Translating experience from health to development

Considering these experiences from the health care sector suggests a number of key considerations and questions about how this approach and method might be translated to or adapted for use in development.

First - in the health sector, the *overall* measure of outcome is the QALY. Generic outcome measures such as the EQ-5D are used to capture the multi-dimensional nature of quality of life and then combined with length of life to estimate QALYs. The QALY is, in effect, itself a bi-dimensional construct: quantity and quality of life. This suggests two possibilities for an approach that might be taken in development:

(a) Treat the overall outcomes from development as essentially bi-dimensional – length of life, adjusted for the development-related quality of life in which life years are experienced (a development-related quality of life, or DR-QALY). A generic development-related outcomes measure could be developed to capture the principal dimensions of quality of life or wellbeing considered important by local populations who are recipients of development initiatives, and which are likely to be affected by development. These would then be used to weight the length of life, in the same way as quality of life is used to weight length of life in QALYs.

However, the argument for treating length of life separately from other dimensions of development-related outcomes is less clear-cut in development than in health, and in fact may not be relevant to many development-related improvements. This suggests an alternative approach:

(b) Treat length of life as one of the set of multiple dimensions, each of which is to be included in a generic outcome measure. This relaxes the requirement that the overall measure of outcome be presented as length of life, weighted for the (development-related) quality of life in which it is experienced. Instead, the overall measure of outcome would be a total score, aggregated from scores on each of the multiple dimensions, where those scores reflect the relative importance of each one. The methods which might be used to achieve that aggregate score are discussed in section 6.2 below. This approach has the merit of not according any special place to length of life, and allowing the outcome measure to directly incorporate preferences about the importance of improvements in life expectancy relative to other dimensions of wellbeing.

Second, just as the development of health outcomes measures rely on various underlying theories of health or utility, devising a generic measure of development outcomes will similarly require identifying a relevant theoretical perspective and clear definitions of the relevant concepts. The literature on Wellbeing in Development (WeD) or from other participatory exercises taken in poor communities would be possible starting points for the development of these definitions. It is striking that participatory exercises of different types reveal a number of strong similarities in the way that people understand poverty and development in different countries and contexts. The importance of physical health and strength is often key, for example, as is the emphasis on physical safety and security. A strong desire for economic security also emerges as important, as do good community relationships. This suggests that the sort of methods described above for defining the 'items' to be

included in a health outcomes measure could provide useful and generally applicable results if applied to development.

This in turn should guide the process to be adopted for item identification and instrument refinement and testing. For example, the approach to be taken may differ depending on whether the goal is to develop a single generic outcome measure which facilitates comparisons between countries as well as within countries. An internationally-applicable 'standard' generic measure of development outcomes would have considerable advantages (e.g., in comparing evidence across countries; and in helping donors to understand how priorities compare across different countries). However, there may be methodological challenges in developing a single standard international generic measure of development outcomes: the process of instrument development would need to seek input from people from a broad range of countries. Persons in different countries may have very different views and cultural perspectives about the dimensions of development which it is relevant to include in such an outcomes measure. To some extent, these differences may be reflected in country specific weights that will be applied to the outcomes in the measure – e.g., the people of one country may place a higher weight on autonomy than on security, and this could be reflected in the local sets of weights applied to observed outcomes. However, the issue is more complex if very profound differences in cultural perspectives affect what items are considered relevant in such an instrument and/or question and scale interpretation.[12] In addition, of course both cultural norms and values may change over time within societies.

Finally, an important consideration for the development of a standard international instrument to measure development outcomes is to ensure that translation of the instrument is undertaken to appropriate standards. This generally entails a forward and backwards translations process, to check that translation has not fundamentally altered the meaning and interpretation of any items. An additional consideration is the development of clear procedures for the mode of administration to be used where it may not be possible to ask local participants to self complete the instrument eg. because of low literacy, or physical limitations.

5.2 Assigning weights and values to outcomes

The instrument development process described above could yield a concise questionnaire, suitable for use with local populations, with items on each of the key dimensions or outcomes found to be relevant to measuring the effects of development. There would be many potential uses of such an instrument, but two are of particular interest. First, the instrument could serve to measure outcomes in a manner that would allow comparison of the effectiveness of alternative development options, and their benefits in terms of improved outcomes relative to their costs. This is comparable to the use of the EQ-5D by NICE, alongside life expectancy, in assessing the cost effectiveness of new health care technologies, and would be highly relevant to the current emphasis on results and value for money in DFID and other donor agencies.

[12] Note that the health measurement instrument, the EQ-5D, is exactly the same in every country, as it is a standardised instrument. But the weights/values on its dimensions/levels are very different between different countries.

Second, the instrument could be used to compare and better understand the importance different people place on different sorts of improvements in development outcomes – e.g., donors compared with local populations. Both these applications require that efforts to develop and use the instrument are accompanied by efforts to assess the weight, or value, of each 'state' defined by the outcomes instrument. A single number is required to summarise each of the outcomes described, which should reflect the relative importance of the various dimensions and levels from some agreed perspective(s).

A range of methods are used in the valuation of health states. The EQ-5D, for example, is accompanied by 'value sets' which detail, for each country where that research has been undertaken, a single index score for each of the 243 states it describes (Szende et al 2007). As noted above, those scores are on a scale anchored at 1 (full health) and 0 (dead) (with values < 0 indicating states worse than dead). This scaling facilitates using the measure as quality of life weights applied to life years in the estimation of QALYs.

Methods commonly used in valuation of EQ-5D states include Time Trade Off (TTO), Visual Analogue Scale (VAS) and Discrete Choice Experiments (DCE). These methods are a subset of a wider set of methods for obtaining 'stated preferences' – that is, participants are asked to imagine living in hypothetical EQ-5D states, and these methods provide a structured way of eliciting people's underlying preferences with respect to the various dimensions and levels described by the instrument. In the UK, weights are derived from a 'large UK population study'. In other words, the relative values of the 243 health states are provided by the general public, not, for example, patients (except insofar members of the public have been, are or will be patients) or, doctors, or (even!) health economists.

Stated preference methods vary considerably in terms of the nature of the tasks which are involved, the properties of the valuations which they produce, and their interpretation. In each case, participants are presented with a series of 'states' associated with various levels on the dimensions included in the outcomes instrument, and then asked to imagine what it would be like to live in that state.

For example, in a discrete choice valuation task, participants are asked to consider two health states (which they may or may not have experienced themselves before) each of which is a different combination of levels and dimensions. See Figure 2. The valuation task itself simply involves saying which of these two states is better. This exercise is repeated for different pairs of states, and the responses can be used to identify the relative importance of the various dimensions to the participants choices.

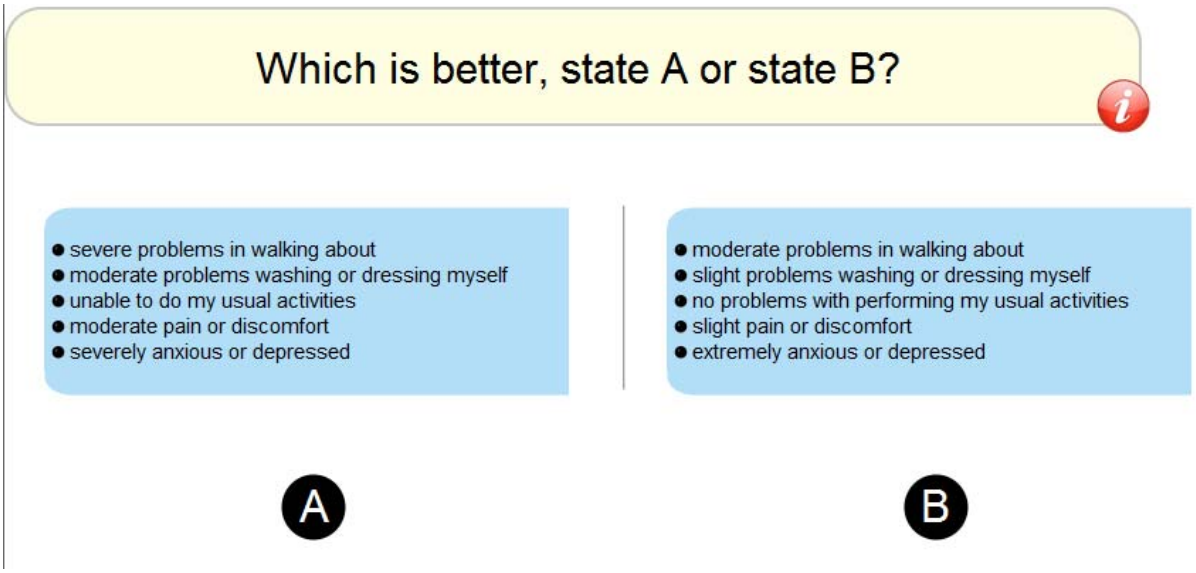


Figure 2 – An example of a discrete choice task involving two EQ-5D-5L states

Source: Krabbe et al (2012).

Visual analogue scale (VAS) valuation asks participants to indicate where they would place the state on a scale marked from 0 (worst possible state) and 100 (best possible state). Discrete choice experiments (DCE) involve the participant comparing two states side by side, and stating which is best. Time Trade Off (TTO) methods ask the participant to imagine living in each state for some fixed period of time (t) and then establishing how many years of life they would be willing to give up to live in a 'perfect' state. Valuation tasks that involve some sort of trade off, such as TTO, are often argued to be superior to methods that require people to directly indicate their values (such as VAS).

All such methods tend to have quite high demands in terms of the verbal and numerical skills required of participants, and even the simplest tasks, such as VAS and DCE, can impose a considerable cognitive burden. While these methods have been widely used in the valuation of health states in western economies, there is less experience with the valuation of health states in poor countries. For example, Jelsma et al (2003) report on a TTO valuation study for EQ-5D in Zimbabwe, but the sample was restricted to urban dwellers, as they were more likely to have the literacy and numeracy required to participate. Methods that restrict the ability to involve samples of people that are representative of local populations have obvious drawbacks as a basis for decision making. That is the case in health state valuation – but such limitations are arguably even more important where the purpose is to understand the priorities and preferences of local people in poor countries who are the target of development initiatives.

This might suggest that the valuation of a generic development outcomes measure will require creating new or adapted approaches which are practical to implement in a wide range of settings and which are appropriate for use with participants from a wide range of backgrounds. For this, researchers could look to the range of participatory research methods developed at the Institute of Development Studies at the University of Sussex, and at the experiences of the Wellbeing in Development group at the University of Bath.

Options might include more reliance on verbal discussion and a mix of qualitative and quantitative approaches. Further, it may be useful to explore options which involve working with groups (eg focus group discussions with people from local communities) rather than relying exclusively on individual-level interviews. Outside the development sector, the set of methods which are commonly used in multi-criteria decision analysis (MCDA) (Devlin and Sussex 2011) might provide a useful starting point. Multi criteria decision analysis is a set of techniques, widely used in public and private sector settings, to aid decisions that are based on more than one criteria. There are a very wide range of specific MCDA methods available, each of which aims to establish the relative weight to be placed on each of the criteria under consideration. For example, the set of techniques that are used to support 'decision conferencing' (Phillips and Bana e Costa 2005) - whereby group discussions and deliberative processes are used to identify and reflect on decisions involving multiple criteria, and used to draw out the weights implicit in trade-offs observed to be made - could be adapted to provide a method for capturing the weights assigned to the various dimensions of outcomes.

6 How could these approaches be used in development?

The potential use in the development sector of a generic outcomes measure derived from the preferences of local populations is enormous. It is likely that the uses of an instrument would evolve with the instrument itself, as was to some extent the case with health care. But a generic outcomes measure could make a useful contribution to a number of current debates and preoccupations in the development sector.

6.1 Use of a generic outcomes measure in RCTs

An outcome measure would not be the first methodology imported from the health sector to development. In recent years much attention has focused on the use of randomised evaluations to measure the impact of development interventions, tracking changes in a population receiving the intervention and comparing with a control group (e.g., see Banerjee and Duflo 2011, Karlan and Appel 2011). Thus far, these evaluations have focused on the benefits of the project in terms of the specific sector and objectives within which they are operating – so, for example, the extent to which microcredit projects increase income or the extent to which the provision of school books improves educational outcomes. This has led to huge improvements in the understanding of development projects.

However, there is as yet no way of comparing the impact of projects across different sectors, or even within the same sector for projects with slightly different objectives. A generic measure, perhaps combined with outcome measures specific to the sector in question, could allow donors and others to compare the value for money of interventions across different sectors using the RCT methodology. RCTs in health, used to evaluate the effectiveness of drugs and other interventions, use generic PROMs such as the EQ-5D in order to make the results comparable and for NICE to use the results in cost effectiveness evaluations

6.2 Use of the outcomes measure in monitoring at the population level

As well as comparing the impact of individual development interventions to each other, an outcomes measure could be used to compare changes in the whole population over time. Using an outcomes measure weighted according to the preferences of poor people would allow an assessment of the extent to which people's lives were improving over time. It could also feed into the development of global outcomes measures such as the debates over a successor framework to the MDGs, and provide one way of ensuring that these debates are informed by the views and preferences of poor people.

Outcomes measures designed for use in a single sector, analogous to the treatment-specific measures in use in the health service, would provide an additional way of monitoring the impact of development projects or policies, and could be used to compare the impact of different interventions in a single sector in a more detailed manner to evaluate their impact and value for money.

6.3 Use in practical decision making exercises

As described above, in the health sector outcomes measures are used as an input into resource allocation decisions, through the NICE decision making process. In development, a measure could also input into donors' and other decision making, and provide a way of ensuring that the views and preferences of poor people fed directly into assessments of value for money in development spending. Describing the outcomes of options under consideration in a standardised way, with the information about the different weights attached to particular outcomes by distinct groups provides a basis for discussion and debate about the relative importance of each aspect of outcomes and thus of different possible uses of resources.

As described above, outcomes measures that were specific to a particular sector, such as agriculture or education, could also enable a more detailed examination of the impact of spending on different projects within a sector, and in turn inform future allocations of resources between projects or programmes.

6.4 Understanding differences in what is valued by different groups

Having a standard set of development outcomes and a methodology for exploring the weights attached to them by the population as a whole would allow comparisons to be made between the priorities of different groups. The different priorities of men and women, for example, could be systematically compared and assessed, and the weights attached to the different outcomes by local communities and by donors, or by communities and government officials, could also be calculated separately and compared. This would allow dialogues about the meaning of 'value' in the value for money discussion to happen on the basis of real evidence about what value in development means for different groups.

7 Conclusion

Development, like any other area of social policy, has to be about making people's lives better. A wealth of research in the sector has shown us how poor people might define 'better', and demonstrated how this differs from what decision makers – governments, official donors or NGOs – might consider to be indicators of improvement. The results agenda provides an extra urgency to ensuring that 'results' are defined in a way which makes sense to those people for whose benefit resources are intended to be used.

Unlike other sectors, however, development research has not made the final push towards embedding knowledge about beneficiaries' values and priorities into decision making. For this, recent innovations in the health sector provide a possible way forward. This paper has described the use of outcome measures in the health sector that are defined and valued according to people's own judgements about the relative importance of different outcomes. We have shown how this approach might be translated across to development, and some of the uses to which an outcomes measure might be put.

The next step in this work will be to test the hypothesis that such a measure could be derived in development through a pilot study with a particular population, and simulations of the use of the measure with policy makers and NGOs.

8 References

Alkire, Sabina (2007), The Missing Dimensions of Poverty Data, Oxford Development Studies 35:4, P. 347-359.

Alkire, S., & Santos, M. (2010). Acute Multidimensional Poverty: a new index for developing countries. *OPHI working paper 38* . Oxford Poverty and Human Development Initiative.

Banerjee, A. and Duflo, E. (2011), Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty, PublicAffairs.

Barder, O. (2009, October). Beyond Planning: markets and networks for better aid. *CGD working paper 185* . Washington DC: Centre for Global Development.

Berwick DM. (1997) Medical associations: guilds or leaders. British Medical Journal (314): 1564 (http://www.bmj.com/cgi/content/full/314/7094/1564)

Booysen, F., Berg, S. v., Burger, R., Malitz, M. v., & Rand, G. d. (2005, August 29-31). Using an asset index to assess trends in poverty in seven Sub-Saharan African countries. *Paper presented at conference on Multidimensional Poverty, hosted by the International Poverty Centre of the UNDP . Brasilia.*

Bowling, A. (2001) Measuring disease: A review of disease-specific quality of life measurement scales. 2nd Edition Ballmoor, UK, Open University Press.

Bowling, Ann (2005) Measuring health : a *review of quality of life measurement scales.* 3rd ed. Maidenhead, U.K. : Open University Press.

Brazier J, Akehurst R, et al (2005) Should patients have a greater role in valuing health states? Applied Health Economics and Helath Policy 4(4): 201-208.

Brazier J, Dixon S, Ratcliffe J. (2009) The role of patient preferences in cost effectiveness analysis: a conflict of values? Pharmacoeconomics 27(9):705-12.

Brazier J and Longworth, L (2011) NICE DSU Technical Support Document 8: An introduction to the measurement and valuation of health for NICE submissions report by the Decision Support Unit, http://www.nicedsu.org.uk/TSD8%20Introduction%20to%20MVH_final.pdf

Brazier J, Rowen D (2011) NICE DSU Technical Support Document 11: Alternatives to EQ-5D for generating health state utility values, http://www.nicedsu.org.uk/TSD11%20Alternatives%20to%20EQ-5D_final.pdf

Brook RH, Ware JE, Davies-Avery A, Stewart AL, Donald CA, Rogers WH, Williams KN, Johnston SA. (1979) Overview of adult health status measures fielded in RAND's Health Insurance Study. Med Care; 17(7,Special Supplement):1-131.

Brooks R (1996) EuroQol: The current state of play. Health Policy 337;1 53-72

Browne et al (2007) Patient Reported Outcome Measures (PROMs) in Elective Surgery. Report to the Department of Health

Chambers, Robert (1983), Rural Development: Putting the Last First, Longmans.
Chambers, Robert (1993), Challenging the Professions: Frontiers for Rural Development, ITDG, London.

Chambers, Robert (1997). Whose Reality Counts? Putting the First Last Intermediate Technology Publications, London.

Chambers, R. (2007, December). Who Counts? the quiet revolution of participation and numbers. *IDS* Working Paper 296 . Institute of Development Studies, University of Sussex.

Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about total knee replacement. J Bone Joint Surg Br 1998;80:63–9.

Department of Health (2008) High quality care for all: NHS Next Stage Review final report. London: The Department.

Devlin N and Appleby J (2010) Getting the most out of PROMs: putting health outcomes at the heart of the NHS. London: King's Fund/OHE.

Devlin N, Dakin H, Rice N, Parkin D, O'Neill P. (2011) The influence of cost-effectiveness and other factors on NICE decisions. Paper presented at HESG, University of York, January 2011.

Devlin N & Sussex J. (2011) Incorporating multiptle criteria in HTA: methods and processes. London: Office of Health Economics.

Dolan, P., Layard, R., & Metcalfe, R. (2011, March). Measuring Subjective Wellbeing for Public Policy: Recommendations on Measures. *Special Paper No. 23* . London: Centre for Economic Performance, LSE.

Ferreira PL, Ferreira LN, Pereira LN (2008) How consistent are health utility values? Quality of Life Research Volume 17, Number 7, 1031-1042, DOI: 10.1007/s11136-008-9368-8.

Filmer, D. & Pritchett, L., 1998. Estimating wealth effects without expenditure data – or tears: An application to educational enrolments in states of India. World Bank Policy Research Working Paper No. 1994, Washington DC: World Bank.

Fukuda-Parr, S. (2010). Reducing Inequality - The Missing MDG: A content review of PRSPs and bilateral donor policy statements. *IDS Bulletin 41,1* , 26-35.

Gough, I., & McGregor, J. (2077). *Wellbeing in Developing Countries: From Theory to Research.* Cambridge, UK: Cambridge University Press.

GPonline (2011) GP leaders fear impact of patient reported outcome measures

Hawthorne, G., Richardson, J., and Osborne, R. (1999). The Assessment of Quality of Life (AQoL) instrument: a psychometric measure of health related quality of life. Qual. Life Res. 8, 209–224.

HESonline (2011) PROMs data
http://www.hesonline.nhs.uk/Ease/servlet/ContentServer?siteID=1937&categoryID=1295

http://www.gponline.com/News/article/1093507/GP-leaders-fear-impact-patient-reported-outcome-measures/

Horsman, J., Furlong, W., et al. (2003). The health utilities index (HUI): Concepts, measurement properties and applications. Health and Quality of Life Outcomes,1, 54.

Hulme, D. (2010). Lessons from the making of the MDGs: Human develompent meets results based management in an unfair world. *IDS Bulletin, Vol. 41, No.1* , 15-25.

Jelsma J et al (2003) How do Zimbabweans value health states? Population Health Metrics 1:11.
http://www.pophealthmetrics.com/content/1/1/11

Karlan, D and Appel, J (2011), More Than Good Intentions: How a New Economics Is Helping to Solve Global Poverty, Dutton.

Kind P (1990) Measuring valuations for health states: A survey of patients in general practice. Discussion Paper 76, Centre for Health Economics, University of York, York.
http://www.york.ac.uk/media/che/documents/papers/discussionpapers/CHE%20Discussion%20Paper%2076.pdf

Kind P, Brooks R, Rabin R (2005) EQ-5D concepts and methods: a developmental history. Springer.

Krabbe P, Devlin N, Oppe M, Kind P, van Hout B (2012) Protocol to value the EQ-5D-5L. Rotterdam: The EuroQol Group.

McGregor, J., Camfield, L., & Woodcock, A. (2009). Needs, Wants and Goals: wellbeing, quality of life and public policy. *Applied Research in Quality of Life, vol.4, No.2* , 135-154.

McKenna S (2011) Measuring patient reported outcomes: moving beyond misplaced common sene to hard science. BMC medicine 9: 86.

Mitchell, A. (2011, June 8). Results for Change. *Speech at the Royal College of Pathologists* . London.

Morris S, Devlin N, Parkin D (2007) Economic analysis in health care. Wiley.

Narayan, D. a. (2000). *Voices of the Poor: Can anyone hear us?* New York: Oxford University Press/World Bank.

National Institute for Health and Clinical Excellence. (2008) *Social value judgements: Principles for the development of NICE guidance.* Second Edition. Available at: http://www.nice.org.uk/media/C18/30/SVJ2PUBLICATION2008.pdf.

NHS Choices (2011) http://www.nhs.uk/Pages/HomePage.aspx

Nightingale F (1863). *Notes on hospitals*, 3rd Edition. London: Longmans.

Norton, A. et al. (2001). *A Rough Guide to PPAs.* London: Overseas Development Institute.

Patrick D, Erickson P. (1993) Health status and health policy: quality of life in health care evaluation and resource allocation. New York: Oxford University Press.

Phillips L, Bana e Costa C (2005) Transparent budgeting, prioritisation and resource allocation with MCDA and decision conferencing, LSE Operations Research Working Paper 05/75. http://www.pophealthmetrics.com/content/1/1/11

Ravallion, M. (2010, November). Troubling tradeoffs in the Human Developent Index. *Policy Rese*arch Working Paper 5484 . Washington DC: World Bank.

Rawlins M, Barnett D, Stevens A. (2010) Pharmacoeconomics: NICE's approach to decision making. British Journal of Clinical Pharmacology 70(3) 346-349.

Rosser R Kind P. 1978 A scale of valuations of states of illness: Is there a social consensus? Int J Epidemiology, 7: 347-358.

Sahn, D.E. & Stifel, D.C., 2000. Poverty comparisons over time and across countries in Africa. World Development 28(12): 2123-215

Seers, D. (1967). The Meaning of Development. *IDS Communication 44* . Brighton, UK: Institute of Development Studies.

Sen, A. (1999). *Development as Freedom.* Oxford: Oxford University Press.

Shah, R. (2011, January 19). The Modern Development Enterprise. *Speech at Centre for Global Development* . Washington DC.

Streiner D, Norman G (2008) *Health measurement scales/ A practical guide to their development and use,* Oxford University Press, (4th edition).

Szende A, Oppe M, Devlin N. (2007) EQ-5D value sets. Inventory, comparative review and user guide. Springer.

Tortora, R. (2009). Sub-Saharan Africans Rank the Millenium Development Goals (MDGs). Washington DC: Gallup World Poll.

UNDP. (1990). Concept and Measurement of Human Development. *Human Development Report* . New York: UNDP.

Van Dalen H, Williams A, Gudex C. (1994) Lay people's evaluations of health: are there variations between different sub-groups? Journal of Epidemiology and Community Health 48:248-253.

Wailloo A, Davis S, Tosh J. (2010) The incorporation of health benefits in CUA using EQ-5D. NICE decision support unit (DSU), available at: http://www.nicedsu.org.uk/PDFs%20of%20reports/DSU%20EQ5D%20final%20report%20-%20submitted.pdf

WeD. (2006, January). *QoL Toolbox*. Retrieved September 9, 2011, from Wellbeing in Developing countries research centrre, University of Bath: http://www.welldev.org.uk/research/methods-toobox/qol-toolbox.htm

White, S. (2009, August). Wellbeing in development practice. *WeD Working Paper 09/50* . Wellbeing in Developing Countries Research Group, University of Bath.

Williams A. (1990) EuroQol: a new facility for the measurement of health-related quality of life. Health Policy 6(3)199-203.

World Bank (2008), World Development Report: Agriculture for Development, Washington, DC: World Bank.

9 Appendix 1 – Sample EQ-5D Questionnaire (EQ-5D-3L)

By placing a tick in one box in each group below, please indicate which statements best describe your own health state today.

Mobility

I have no problems in walking about ❑

I have some problems in walking about ❑

I am confined to bed ❑

Self-Care

I have no problems with self-care ❑

I have some problems washing or dressing myself ❑

I am unable to wash or dress myself ❑

Usual Activities (*e.g. work, study, housework, family or leisure activities)*

I have no problems with performing my usual activities ❑

I have some problems with performing my usual activities ❑

I am unable to perform my usual activities ❑

Pain/Discomfort

I have no pain or discomfort ❑

I have moderate pain or discomfort ❑

I have extreme pain or discomfort ❑

Anxiety/Depression

I am not anxious or depressed ❑

I am moderately anxious or depressed ❑

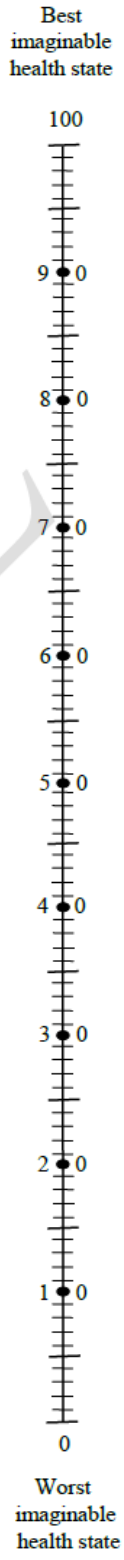
I am extremely anxious or depressed ❑

SAMPLE

To help people say how good or bad a health state is, we have drawn a scale (rather like a thermometer) on which the best state you can imagine is marked 100 and the worst state you can imagine is marked 0.

We would like you to indicate on this scale how good or bad your own health is today, in your opinion. Please do this by drawing a line from the box below to whichever point on the scale indicates how good or bad your health state is today.

Your own health state today



SAMPLE

Source: http://www.euroqol.org/eq-5d/how-to-obtain-eq-5d.html

[i] The Oxford Hips Score is an example of a condition-specific instrument (Dawson et al 1998). It has 12 items to assess symptoms and functional status (disability), with each item having 5 possible answers. The summary score for a health state described by OHS is obtained by adding the levels of each item resulting in a score between 12 and 60, where 12 is the best outcome.