GLOBAL EDUCATION LEADERS' PROGRAM

New Metrics Briefing 2: Feedback to the GELP Metrics co-design group

Simon Breakspear, Cambridge University

The paper was prepared by Simon for the new metrics co-design group. We share it with the community as an insightful framing of a central element in the push for new metrics: moving the evaluation of education to a focus on learning. We hope that the discussion prompts in the second half of the paper will be taken up in our new metrics sessions during the event.

Much of my discussion focuses on the OUTPUTS element of the logic model, as I feel that this is the key 'lagging factor' in our assessment and indicator capabilities. I structure the feedback in two steps. First, I outline three key framing ideas to help ground the discussion of metrics on student growth and, learning and teaching. Second, I describe potential challenges for both the strategic use of current metrics and the development of new metrics. I hope these may be useful prompts for debate and discussion.

Three key Framing ideas

1. Connect the GELP new metrics discussion to the wider literature called for the reimagining of the purposes and design of educational assessment.

It is possible that a focus on 'metrics' jumps too quickly over the concepts and debates currently occurring in the field about reforming the design and purposes of assessment in education systems. The idea of defining new metrics can give the impression that we can talk about aggregated indicators of performance, equity and efficiency at the system level, without dealing with the fundamental challenges of assessment and measurement.

This year alone three major reports have been released that engage with the issues of reforming educational assessment. Whilst, they do not go far enough into the education 3.0 agenda, they do highlight a growing dissatisfaction with current practices and their effects on actors' (students, teachers, administrators, political leaders) across the system. I recommend delegates engage with at least the executive summaries of these reports.

- Masters, Geoff N., "Reforming Educational Assessment: Imperatives, principles and challenges" (2013).
  The Gordon Common on the Future of Assessment in Education (2013)
- OECD, Synergies for Better Learning: An international Perspective on Evaluation and Assessment (2013)

2. Bring to the fore a focus on the use of metrics to improve teaching and learning, rather than accountability and certification

Connecting the focus, purpose and design of metrics is crucial. I endorse fully the discussion in the workshop overview that explores a framework on the purposes of new metrics and how this should align with the design features. My addition to the current outline is to propose that metrics should focus on the core purpose of seeking to continuously measure, assess and improve teaching and learning at all levels of the system. I believe this is at the heart of new learning systems, and should be the focus of the 'new strategies' for using existing and new metrics.

There are three key purposes of educational assessment in the literature:

1. Certification and selection (sometimes treated separately).
Certification typically is carried out at the completion of high school, although in many systems, including the UK, it is a two-step process (GCSE and A-Level). It is important to understand that certification and selection uses of metrics are high-stakes for students, and thus attract substantial parental and media attention. There is typically a high political risk to making changes to these.

2. Accountability
Another long-standing use of assessment and one that has gained great prominence in recent years is for the purpose of holding providers (systems, schools and teachers) directly accountable for the performance of their students.

When any metric is used for high-stakes activities it often leads to gaming behaviours by the key actors under pressure – whether they be students, teachers administrators or even policy-makers. The best documented case of the potential consequences of high-stakes test-based accountabilities programs is the US No Child Left Behind efforts which lead to wide-scale teaching to the test, narrowing teaching and learning and at times cheating and corruption.

3. Improving teaching and learning
The core role of assessment should be to make judgments about the extent of student growth and as a consequence provide feedback on teaching and learning decisions. Typically assessment for improving teaching and learning has been categorised as 'formative assessment' and is thought to be an activity of teachers embedded in classroom practice. This has meant that school and system level metrics have typically adopted a focus on certification and accountability purposes, but without building any necessary loop back to improve the core of education – teaching and learning. This, I believe is a fundamental error. PISA, for example is not designed to improve teaching and learning, as a system assessment it tests a random sample of students across a random sample of schools, and thus has no method of connecting outcomes with instruction.

I like Masters' idea that metrics at different levels of the system should focus on decision making to improve learning:
"the use of assessment to inform decision-making parallels the use of assessment in other professions such as medicine and psychology, where the purpose is not so much to judge as to understand. Professionals use assessments to better understand the presenting situation or problem, to identify starting points for action, to decide on appropriate evidence-based interventions, to monitor progress, and to evaluate the effectiveness of the decisions they make." (2013: 3)

So what do system leaders what to know about the performance of their education systems over time? What ramifications may this have on the design of new metric systems?

3. Don't lose a focus on learners!

Lastly, metrics discussions quickly move towards thinking in terms of systems rather than students. Whilst, aggregated metrics of system performance and progress are important, the fundamental purpose *of assessment is to establish where learners are in their learning at the time of assessment.*
Assessment should be conceptualised as the process of establishing where students are in their long-term learning and what progress they are making over time, usually in terms of their developing knowledge, skills and understandings.

We should first focus on how to make accurate judgments about student growth and development on the new deep learning outcomes. Once we can do this, aggregating to the system level is a simple next steps. In short, all the uses of metrics first depend on good assessment information about where students are in on or more aspects of their learning.

Discussion areas;

'Strategic uses of current metrics'

1. How could there be a rebalancing of the current purposes of outcome metrics?

To what extent are the accountability purposes of current metrics in GELP jurisdictions overpowering the purpose of improving teaching and learning or the ability to move towards deeper learning goals and new pedagogies?

Reflecting on the American context the Gordon commission concluded:
"Conceptions of what it means to educate and to be an educated person are changing. Notions of and demands on practice in the teaching and learning enterprise are broadening and expanding. And the concern with accountability forces this dynamic and eclectic enterprise to constrict and, in the worst of instances, to compromise in the interest of meeting certain accountability criteria. These realities, coupled with changes in epistemology, cognitive and learning sciences, as well as in the pedagogical technologies that inform teaching and learning, are narrowing — possibly even stifling — creativity and flexibility in teaching and learning transactions." (Gordon commission, 2013, p16).

2. How could new analysis of existing metrics bring new insights about system performance and effective teaching and learning?

Education systems are complex and headline data very rarely provides the true story about performance and progress. For example, a new report[1]. by Carnoy and Rothestein "what do international tests really show about student performance?" re-analysed the PISA data. In short they argue and demonstrate that once social class inequality is taken into account the relative performance of US students is much better than is appears when countries' national average performance is conventionally compared.

How could the jurisdictions invest in small high-quality teams of statisticians and educational researcher that could ask new deeper questions of existing data?

3. What will be the responses of different stakeholders to the reform of assessment and metrics?

Shaping the political economy of assessment and metric reform is of central importance. Currently, parents, and other groups, predominantly expect assessment to carry on as they experienced it. It is imperative to involve and engage them in the process of change.

'Technical foci – designing new metrics'

1. How can the new metrics gather evidence about progress on deeper learning goals?
I wholeheartedly agree with the statement of the GELP design group that it is important to, **"**be clear about what their goals are, before moving to identify specific metrics".

Most metrics of interest will be focused on outcome variables rather than inputs. Of particular interest is what students should know, understand and be able to do across a new set of deep learning goals.

The metrics discussion should first be grounded in a detailed discussion about the new 'deeper learning goals' and aims for education systems in the 21st century. In the first GELP metrics paper I offered the National Research council's framework of cognitive, interpersonal and intrapersonal competencies. Whilst, this generic framework is useful as a starting point, the work must now be put into conceptualising these broad definitions as conceptual frameworks – explicitly laying out what students should know, understand and be able to do.

As an example in the cognitive domain, the PISA assessment frameworks for mathematics, reading and science. The frameworks seek to operationalise the knowledge and skills that are to be assessed in each domain. Each area framework identifies both the knowledge of fundamental skills and understandings required and then also the capacity to use those skills to address real-life problems.

The first step in the development of new assessments and metrics will be to generate comprehensive conceptual frameworks and development progressions across each of the target outcome domains.

2. Do they need to be internationally comparative metrics?

Internationally comparability requires that all countries use the same target outcomes, instruments and scale. That is we compare along one single dimension. Problems of comparability have been a central theme of critique of PISA, TIMS and PIRLS. The core challenge is how to ensure comparability of meaning for assessment scores across a range of countries, cultures, languages in addition to the diversity of education system aims and curricula. It is difficult enough to generate comparable measure across a domain such as reading. How much more therefore is areas such as interpersonal skills, entrepreneurship and resilience? Any challenge to the capacity to make comparable judgments between countries brings doubt to the validity of the indicators generated.

PISA is phenomenally attractive as it simplifies the complexity of the world's education systems into three key macro-indicators (reading, mathematics and science). Whilst this is attractive (and almost seductive) in its simplicity, serious questions should be raised as to what is actually being measured. One scholar has argued concludes that, *"PISA will*

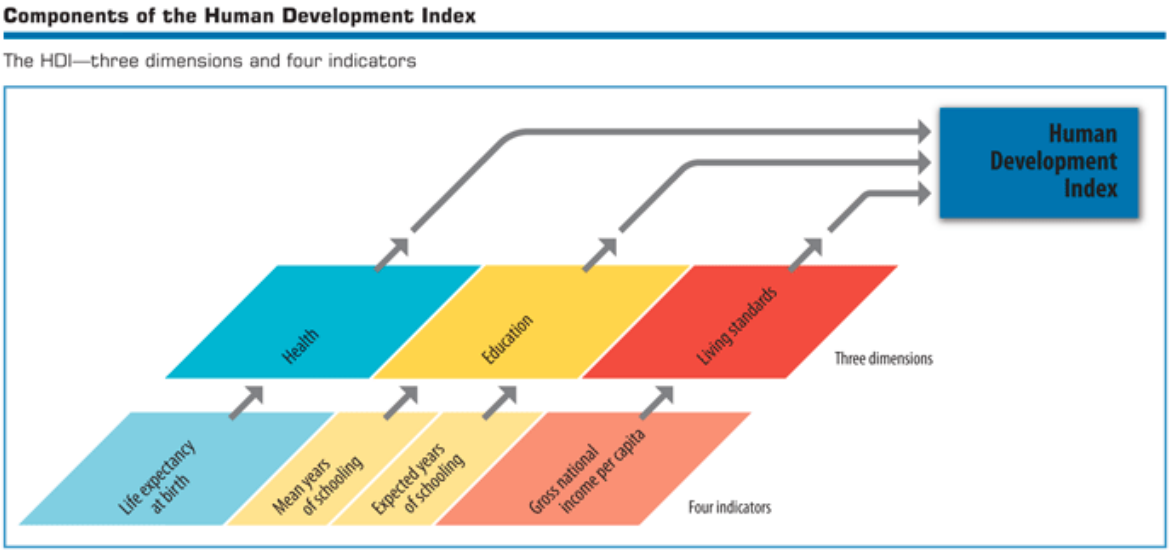
[1] http://www.epi.org/publication/us-student-performance-testing/

continue to measure not so much the outcome of education systems as the historical and cultural contexts of countries" (Bonnet, 2002).

3. Single macro-indicators/ indexes or gather evidence across multiple dimensions

Increasingly the best approach to thinking about metrics is the collection of multiple streams of evidence in order to make judgments about whether or not learning is improving.

Even when we do have multiple dimensions then the tendency is to still combine them together into an aggregated indicators or metrics for the purpose of ranking and rating. Take for example the UN's Human Develop Index. This was no doubt a substantial step forward from discussion of GNP per capita as a proxy for human development[2] (see figure below). The HDI aggregates health, education and living standards into one index and then rates countries globally. The aggregated measure is very appealing. But in order to have any understanding of what is really going on, then you need to go below the surface again. Countries with similar UN HDI scores may vary substantially across the four dimensions, e.g. one being higher in health outcomes, and low in education, whilst the other may have the opposite strengths and weaknesses. Saying that these two countries are the 'same', is utterly inaccurate. In one, you are unlikely to get a good education. In the other you are likely to die young.



Policy makers likes aggregated macro-indicators that rank systems or sub-elements of systems (states, schools, classroom) along one scale. This is particularly true for areas where accountability and changing actors behaviour is central to the desired end. But if improving system performance is the driving purpose, then keeping multi-dimensionality and avoiding simplistic aggregation is the approach that should be taken.

4. Driving R & D for new assessments

The future of assessment and metric will be influenced by the R & D activities that can be stimulated. How can GELP stimulate a new supply of next generation assessments (by communicating a clear global demand) that will allow the formulation of output indicators that reflect the desired deeper learning goals of systems moving towards 3.0? Who might be the key partners in this work? (ETS, OECD, IEA, ACER, Pearson).

5. What will Big Data change?

At the moment we construct artificial scenarios called 'tests' to work out what students can and can't do – but instead, the digital revolution allows us to use naturally occurring data to build a richer and more realistic picture of what a child can do. This 'evidence centred design', as the authors of *Digital Ocean* have called it, could break us away from the problems of

[2] http://hdr.undp.org/en/statistics/hdi/

high-stakes assessment as it stands. A broader range of activities can be used to form judgments of students, and these can become simply 'work products' – in other words, what a child is producing on a daily basis in response to their lessons.

The driving question must be: What are the challenges to current assessment design, delivery and analysis that digital data will help us solve?

Further reading on Big Data:

DeCerbo and Behrens**.** *Implications of the digital ocean on current and future assessment.*

Mislevy, R., Behrens, J. T., Dicerbo, K. E., & Levy, R. Design and discovery in educational assessment: evidence-centred design, psychometrics, and educational data mining. Journal of Educational Data Mining, 4(1), 11–48.