RISE

RESEARCH ON IMPROVING SYSTEMS OF EDUCATION

WORKING PAPER SERIES

November 2016

The good, the bad, and the ugly – testing as a part of the education ecosystem

Newman Burdett

RISE-WP-16/010

November 2016

Funded by:

The good, the bad, and the ugly – testing as a key part of the education ecosystem

Newman Burdett

November 2016

Abstract

To improve the quality of education systems, we need good information about learning outcomes to guide and inform policy decisions. Without good, accurate information, logical decisions can be made that can be disastrously wrong. Unfortunately, the world of education has proven this point on many occasions. This RISE working paper will highlight this issue in relation to testing, assessment, and measurement.

In order to answer the question, “What works to improve education systems to deliver better learning for all at scale in developing countries?” we need to first acknowledge that in many of the education systems in these countries, there is either no effective monitoring of learning or the current testing regimes are part of the problem.

This paper will attempt to address the issues surrounding educational measurement and how this plays out in ways that are relevant to education systems that are struggling to improve quality of learning. This is not just a case of gaining good information to guide and evaluate reform in these systems; the assessments need to be considered an integral part of the system. The evidence shows that examinations and tests are powerful drivers of behaviour and need to be closely aligned with the desired education outcomes to achieve good quality learning. Understanding how poorly designed assessment and accountability systems undermine education systems, and conversely how to integrate well-designed measurement within the system, is an important step in improving learning outcomes.

Key Words

education systems, assessment, measurement, quality learning

Introduction

On 1 June 2009, Air France Flight 447, a brand new Airbus A330, took off from Rio de Janeiro on route to Paris. Somewhere in the middle of the ocean, it flew into a thunderstorm and vanished. When the wreckage was finally recovered and analysed, it turned out that sensors, called pitots, in the external skin of the aircraft had become blocked by ice depriving the aircraft and pilot of important information about airspeed and whether the airplane was pointing up or down. In the absence of this information, in the dark and with no external references and no feedback on the results of his actions, the pilot made a series of fatal mistakes pushing the plane into a steeper and steeper climb until it fell out of the sky in an unrecoverable stall.

On the surface, this might appear to have nothing to do with education, but if we consider education as a dynamic system then the message of Flight 447 becomes very relevant and telling. Without good, accurate information people can make logical decisions that can be disastrously wrong and have no idea how to correct them. This applies very much to education systems; and unfortunately, the world of education has many examples to learn from. This working paper will highlight things that go commonly wrong linked to testing, assessment, and measurement, because without good measurement we are metaphorically flying in the dark with no clue.

It is important to stress here that this working paper is not intended to focus on the issues of testing that cause so much controversy in the more economically developed countries although many of the issues, teaching to the test, poorly thought out accountability regimes etc., are similar and many of the examples discussed later are taken from the developed world as a warning not to import policy mistakes. This is not to downplay the often heated debates about the type and amount of testing going on in places such as the UK or the US (and we will look at examples from both countries later) but in order to answer the question, “What works to improve education systems to deliver better learning for all at scale in developing countries?” we need to first acknowledge that in many of these developing countries’ education systems there is either no effective monitoring of learning or that the current testing regimes are part of the problem.

This working paper attempts to highlight how the issues surrounding educational measurement play out in ways that are relevant to education systems that are struggling to improve quality of learning. This is not just a case of gaining good information to guide and evaluate reform in these systems; the assessments need to be considered an integral part of the system. The evidence shows that examinations and tests are powerful drivers of behaviour and need to be closely aligned with the desired education outcomes to achieve good quality learning. Understanding how poorly designed assessment and accountability systems undermine education systems, and conversely how to integrate well-designed measurement within the system, is an important step in improving learning outcomes.

Why is educational measurement important?

Like all complex systems, education relies on a series of feedback loops at all levels of the system to ensure children are effectively learning the right things that they need to succeed and prosper in life. Without good information about how effective the learning is, educational quality will be judged by a series of proxies that might bear very little real relationship to the reading, writing, or mathematical skills of the students – for example a teacher’s effectiveness might be judged purely on seniority or how well they maintain class discipline rather than how effectively the children in their class learn. And if there is a test, and if the test is bad, then the system will be judged by how well it prepares learners for the test rather than for life beyond school.

In many countries, the examination is either the *de jure* or the *de facto* education system, with education defined explicitly in terms of the examination. This means that examination results are often much more important to stakeholders than learning. In these countries the stakes are very high with the examination determining access to jobs, to further education, fields of study, etc. In the more economically developed world, this is often mitigated as there are alternative employment

opportunities, other routes to qualifications, and broader access to universities and higher education. These alternative routes are often lacking in the developing world, making school examinations often the only gateways to opportunities.

If the assessment results are well aligned with the desired learning outcomes, i.e. the assessment measures the skills students really need to have to succeed in life, then this is not a problem. But in many cases, the examination system is corrupted, either by being open to cheating and bribery, by being incompetent and error prone, or by measuring things that are not relevant outside of school e.g. rote learning. In these cases, the whole weight of performance accountability (both from the state and parents) is designed to promote 'bad' education over useful learning.

The waters are further muddied by the complex nature of educational assessment, which can look very simple on the surface, but is a highly technical field with many dimensions. It is therefore good to recap the fundamentals and understand what makes good assessment, before starting to untangle how things go wrong at a system level.

Types of assessment

There is a lot of nuanced terminology surrounding educational measurement some of which, e.g. terms such as 'testing', can be quite charged (the example discussed later on high stakes testing and corruption in the USA partly illustrates why it is such a heated area) so it is important to define our terms. Assessment is the process of gathering and evaluating information about learning (of knowledge, skills, attitudes, beliefs, etc.) in measurable terms. Within this broad term of assessment, there are many specific purposes and activities which go under many names but, broadly speaking, all forms of assessment can be put into three important and useful categories.

At the highest system level there are *monitoring assessments* – these let us know how the system is doing and what the population of students can and cannot do. These are programmes such as national monitoring assessments, PISA, EGRA, etc. They are aimed at producing high quality information at a regional or national level, but not necessarily at producing good data at an individual level. These are vital for measuring the quality of learning and any changes at a system level because neither of the other types of assessment reliably provide this information.

The next broad category includes the end of cycle or school leaving examinations such as end of primary examinations, school certificate examinations, etc. These tend to be certified examinations that provide the student with a proof of education, but they also include entrance examinations and tests linked to accountability. These tend to be high stakes examinations in that they provide the evidence to allow access to further education choices, employment, or funding. They tend to be summative assessments, i.e. testing what has been learnt up to that date and providing a snap-shot of achievement. When this paper refers to *tests* and *examinations* in this article, we mean summative, high stakes assessments.

Finally, and the category that tends to be least well understood and implemented, are the *formative assessments.* These are the day-to-day assessments designed to feedback into learning in a dynamic way, to improve and direct learning. There are many different definitions of formative assessment, but a useful one is from Black and Wiliams (2009), which is broad enough, but also clear enough to be applied meaningfully. They state that "practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded than the decisions they would have taken in the absence of the evidence that was elicited."

It follows from this definition that the intention for formative assessment is to be used in immediate, short-term feedback loops to improve individual student learning – this distinguishes it from the system level assessments which provide feedback at the system level and work on longer term feedback cycles. It also distinguishes formative assessments from end of cycle assessments

aimed at providing evidence *after* the event as a one-off summative statement of what learning has been achieved.

Serious problems can arise if policy or practice combines or confuses these purposes. Similarly, problems also arise if the assessments themselves are not good or not fit for purpose. To avoid these problems, an education system has to have good information and to use that information wisely, and for that to happen, there needs to be a good assessment system.

What makes good assessment?

In the Future of Assessment: 2025 and Beyond[1], I argued that:

"Given the complexity and debate that surrounds education and assessment, nationally and internationally, it is hard to say categorically what 'good assessment' is – values and cultural influences blur the borderlines – but we can state that good educational assessment needs to meet some basic criteria.

Firstly, the assessment needs to have a clearly defined purpose. There is no point assessing needlessly or placing unrealistic or potentially conflicting demands on the assessment. Experience teaches us that there will be counterproductive tensions if a single assessment is used to monitor national standards and act as an accountability measure for teachers. If we do not understand why we are asking the question, we will struggle to understand the answers to that question.

Secondly, it must be fit for that purpose: it must measure what we want learners to learn. It is good that Ofqual[2] is now focusing on the validity of assessment to ensure that the results are meaningful, useful and appropriate rather than just repeatable (Stacey, 2014). Good assessment needs to reflect everything that we consider important to a good education – it is not a case of if important things should be assessed, but how they are assessed. Good assessment should start from the intended learning and does not mean valuing only what we can measure well, but finding ways to measure what we value.

Most importantly, and often most overlooked, good assessment should borrow from medicine the principle of *primum non nocere* – it should do no harm, in this case to the learners."

For the purpose of this paper it is worth unpacking the second point - that assessment must be fit for purpose - in some detail. To understand what makes a bad assessment, it is necessary to understand technically what makes a good assessment. One of the key measures here is reliability – if we repeat the measurement, will we get the same result? Validity is the other linked principle here – does the assessment measure what it claims it is measuring? This is a significant problem with many assessments because reading, writing, and mathematics ability are complex psychological constructs that cannot be directly observed. Unlike a physical variable such as height, which can be measured directly, constructs such as language ability or mathematics ability can only be measured indirectly by setting tasks and judging the response.

It is very easy in practice for these proxy measures to become uncoupled from the underlying construct that an assessment is attempting to measure and end up unintentionally measuring something else instead, e.g. a mathematics problem with a complex introduction might end up being more a measure of a student's language ability than their mathematics ability – it is not the mathematics that the student does not understand but what the question is asking them to do.

This is not a purely philosophical concern, there are examples of bad assessments out there at all levels giving incorrect results and misleading information. For example, statistics for one of the items (questions) reviewed for a National Assessment Survey that was supposed to be assessing

[1] http://filestore.aqa.org.uk/pdf/AQA-THE-FUTURE-OF-ASSESSMENT.PDF
[2] The UK's non-ministerial government department that regulates qualifications, exams and assessments.

whether students understood number position or place value (i.e. do they understand that in a three-digit number one digit represents hundreds, the one to its right represents tens and the final digit represents ones, a basic but important part of mathematical knowledge) indicated that students did not understand this concept at all. However, the item[3] read:

"What is the number in the place value of the units in the cubed root of 531441?"

It is clear that this is not assessing place value, but whether students can accurately find the cube root of a large number – a very different and much harder task. This assessment was saying that students cannot perform a basic function when in reality there is no good information on what they do and do not know about place value. The only useful deduction from this item is that test developers needed more training and to have better quality assurance procedures. More detailed study of the item performance showed it had a very low facility (the chance of an average student getting it right) - at around the same level as guessing - and very low discrimination (i.e. bright students were no more likely to get it right or wrong than weak students).

A good assessment has to actually measure what it says it is measuring, it has to do it reliably and reproducibly, it has to be at an appropriate level and difficulty for the intended students, and it has to be able to discriminate between students who have high ability and students who do not. A significant proportion of the items on this National Assessment Survey were incorrectly measuring student performance: some were factually incorrect, some had no correct answer, some had several correct answers, but only one of which would score, and some were so confusing it was impossible to say what the correct answer was.

The example given above is an extreme case of a bad test but even with less extreme examples, poor reliability, validity, variable standards or even just inadequate design or lack of clarity about purposes, can have more or less damaging consequences for education systems. Producing good assessments is a technically demanding exercise that requires expertise. Even though there are some good institutions in developing countries with skilled and experienced practitioners, this assessment expertise is often limited or lacking. The lack of good information about learning makes it very hard to make good decisions about the quality of an education system.

What happens if assessments give you bad or insufficient information?

Many education systems have suffered some variant on 'PISA shock' (Waldow, 2009) – a phrase coined when Germany first took part in PISA and to the national horror discovered that the collective presumption that Germany had a world-beating, high-performing education system was unfounded. Without good assessment coupled to reliable standards, it is impossible to know exactly how an education system is performing and it is very easy for standards to drift. For example, in the UK, despite having very good records and statistics on the national examinations (the General Certificate of Secondary Education) and a large amount of debate and research, it was impossible to determine whether or not educational standards had improved or declined over several decades. In response, the UK government introduced a National Reference Test at Key Stage 4 (age 16) to specifically monitor standards. The important point here is that it is very difficult to use a school certificate examination for monitoring purposes and *vice versa*.

This unreliability in assessment does not necessarily show at a national level and (even if national statistics are stable and reliable) quite often there can be large uncertainty over individual results leading to misclassification of students, i.e. students getting the wrong grades. This is not only unfair on the individual, but this can have significant system level impacts.

End-of-primary examinations are often used to decide students' secondary education pathways, either by school choice or by directing a student to a specific academic or vocational education stream. But the truth is these examinations are often very poor predictors of future performance. In

[3] The numbers have been altered in case the item is still being used but otherwise the item is the same.

2004, the Seychelles announced a move from a system where the top 20-30% of students were allowed to study for the O Level (a secondary school leaving examination) and the rest had to sit the National Exam at Secondary 5, to a system where all students could sit for the same examination, the International General Certificate of Secondary Education. There was a lot of concern at the time that this would lead to a drop in pass rates and that a drop would cause political upheaval. In the event, pass rates stayed remarkably stable despite the increase in student numbers and the absolute numbers of students getting the higher grades increased. The new examination was a well-respected international qualification with robust standards and good year-on-year comparability and so the best explanation for this increase in students gaining the higher grades was that the primary examination had been misclassifying many ultimately high achieving students as having low potential. This means that, prior to the change in exam, a significant number of students with high potential had been lost to the labour market as higher earning jobs or access to further study required students to have O level certificates. Introducing an examination that actively promoted equity of opportunity allowed the government to achieve some success in their policy aim of providing effective 'education for all'.

These case studies highlight the issues present in most assessment systems and the difficulty of getting it right, but what happens when things go really bad?

What happens if you have a bad test?

The Seychelles was a good example of what happens when you have an examination that is not good at predicting the students likely to achieve, but this section discusses assessments that lack validity, i.e. assessments that are not just bad at reliably selecting the right students, but are selecting students on the wrong criteria.

Nobody would argue that Pakistan's education system faces huge challenges (UNESCO, 2016) with many children, especially girls, failing to even reach education. Even if they manage to be educated, there are serious concerns over the quality of that education. A significant part of the problem is the examination system in Pakistan, and this is true all the way from the primary school examinations through to the middle school and high school examinations. These examinations have many failings, including wide spread corruption and malpractice, poorly set questions, and poor marking and data entry. These failings are well documented but, arguably, the biggest impact on the quality of education is on the reduction of teaching to rote learning and a very narrow curriculum (Kamrani, 2011; Rehmani, 2003; Mirza, 1999). As Rehmani characterises it:

"Exam questions are repeated at least every three to five years and hence questions can be predicted. There are 'model papers', or 'guess paper guides' available in the market with ready-made answers based on past five years' papers. Teachers and students tend to rely on such guides and put their content to memory. Regurgitation seems to be the only key for students to pass the examination rather than creative thinking and independent analyses".

Our own fieldwork also saw evidence of this and the gaping disconnect between what should have been taught and what was being taught. In one notable, but not unique example, a teacher was teaching a comprehension class in English by means of a chant response rote learning exercise. The teacher read from a selected passage from the textbook, stopping at the extracts that are used in the examination questions, which she then repeated, followed by the question and the answer, getting the class to repeat this back and forth, chanting several times before moving on. For example:

> Teacher (reading): … the clouds hung dark and heavy. The clouds hung dark and heavy. What phrase does the writer use to suggest a storm is approaching? The clouds hung dark and heavy. What phrase does the writer use to suggest a storm is approaching? The clouds hung dark and heavy. (aimed at the class) What phrase does the writer use to suggest a storm is approaching?

Class: The clouds hung dark and heavy.
Teacher: Correct – the clouds hung dark and heavy. Asmaa, what phrase does the writer use to suggest a storm is approaching?

Asmaa: They hung dark and heavy.

Teacher: No, THE CLOUDS hung dark and heavy.

This is not comprehension; this is rote learning, a lower order skill according to Bloom's taxonomy. Students are expected to exactly repeat the scoring phrases in the examination.

This teacher worked in a school with a very good reputation, very good facilities, had excellent classroom control, spent all lesson very much on task, and was obviously popular with her students. Yet despite all of this, the quality of comprehension learning was very low. Interviews with the students showed that they were incapable of interacting with the text in any meaningful manner, other than to answer the learnt responses.

This is fundamental to understanding what is meant by the quality of an education system, exemplified in a system such as Pakistan, where the education is determined by the examination and the only measure of success is a rote learning of the answers to a test, and not the quality of learning. The achievement in the examination gives no information as to the real abilities of students. Employers and universities have no means of selecting the best applicants, but instead have to rely on the examination results to inform those choices. Given this, it is no wonder that parents, in their rightful ambition to see the best opportunities for their children, place great value on gaining these examination results, even if they know them to be deeply flawed. One of the sad findings of the field study was the high incidence of parents who withdrew their children from good schools that they felt were providing a good education and delivering real learning (as measured later by alumni students' performance on university entrance tests and ability to succeed once at university). Instead, these parents felt compelled to place their children for the years prior to the examination into a system that the parents felt was inferior in terms of what was learnt, but one that they thought it was easier to gain marks in. The students themselves were deeply resentful of this change and felt this period preparing for a flawed examination was a waste of education and a cessation of real learning. The linking of these bad examinations to high-stakes outcomes (jobs and university entrance) gives these examinations great power that subverts any attempt to improve the system.

If everyone in a society has only experienced the same deficiencies, it is very difficult to judge what is good. In the absence of any meaningful criteria for judging what is good education, marks and certificates have become a proxy for learning and have achieved a higher value than learning itself. As Adnan and Mahmood (2014) summarise it:

"…need for success in HSSC[3] exam has prompted teachers to such teaching where principal objective is to get marks ignoring learning needs of the students… (the) exam has emerged as a leading factor that has motivated some sub-factors such as teachers, students and other stake holders and in turn all of them have given their share in restricting teaching to a question paper."

This is not confined to just the high school examinations, as all the examinations are seen as high stakes to a larger or lesser degree. For the child and parents, they mean progression and access to scholarships or improved chance of entry to the school of choice. For schools, results bring prestige and increase their attractiveness to parents, who in the absence of any other viable indicators can only judge the teaching at the school by examination results.

This replacement of learning by marks and certification is almost universal in Pakistan, and the fact it is highly entrenched can be seen by remarks made by various members of the Pakistan parliament when exposed in a fake degree scandal:

[3] Higher Secondary School Certificate

“A degree is a degree! Whether fake or genuine, it’s a degree! It makes no difference!” Baluchistan province chief minister Nawab Aslam Raisani.[4]

The learning does not matter, just the façade; education reduced to a cargo cult.

Whilst the meaning of what is a ‘good’ education is open to debate, it is clear that these bad examinations are overriding education systems to churn out students who lack the skills needed for employment or further study. Furthermore, they are robbing employers and universities of the ability to easily identify the candidates they wish to attract, leading to poor productivity, high dropout rates, and hampering economic development.

These bad examinations not only have severe and direct impacts, but also often have a subtler, distorting effect on system feed-back - not necessarily bad assessments, but the bad use of assessments.

Feedback loops and subtler effects

Assessments have indirect impacts and these can be detrimental, especially where the assessment results are used inappropriately or carelessly. The world of assessment and education has numerous examples of Campbell’s Law and Goodhart’s Law. Campbell’s Law (Campbell, 1976) says that, "The more any quantitative social indicator (or even some qualitative indicator) is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." Goodhart’s Law (Goodhart, 1984) states that, "When a measure becomes a target, it ceases to be a good measure.”

This is a significant issue for evidence based policy and using student outcomes to drive educational improvement, because poorly thought through incentive schemes that rely on only one measure (or have no external verification of that measure), can be easily perverted to give unexpected consequences.

In the UK, the Department for Education tried a raft of incentive schemes and monitoring schools by results to boost academic performance. They published league tables of numbers of passes to increase competition between schools; to encourage education post-16 they paid further education colleges, sixth-form colleges and school sixth forms by the number of qualifications a student takes. Over time, and against increasing pressure to perform, schools started to evolve tactics that maximised their performance measures, sometimes to the detriment of their students.

Schools were found to be abetting cheating in examinations, some were forcing students to take exams early and retake exams many times to boost results, others removed students at risk of failure from the roster. Others took a very cynically commercial view and chose qualifications to maximise their institution’s profitability.

This finally culminated in a series of reports that found evidence that, “some schools or colleges maximised income by ‘piling up’ young people with low-quality qualifications which did not offer progression.”

“The Wolf Report (Wolf, 2011) demonstrated that the current performance table system creates perverse incentives. Schools have been tempted to teach qualifications which attract the most points in the performance tables - not the qualifications that will support young people to progress.

The current system incentivises schools and colleges to load too many students with low-quality, small or easy qualifications, often in random combinations, that employers do not value (Gibb, 2011).”

[4] Reported in The Dawn 30 June 2010

This is a classic example of these laws coming into play and school leadership trying to maximise performance scores over learning which ends up distorting the system - and in a competitive system, once even a small minority start to cheat and are known to be gaming the system, it puts more pressure on those not gaming the system to improve their performance measures to ensure they have fair access to funding. To use an economics analogy, "bad money drives out good if they exchange for the same price" (Mundell, 1998), i.e. low quality qualifications will drive out high quality ones if they are given the same value.

As Michael Gove, the Secretary for State responsible for education at the time, phrased it, "The school is in effect gaming the system by not thinking what is in the best interest of the student, but using the student as a means of gathering points so that the school itself can look better. When a small minority cheat, the system is corrupted for others. That has to stop."

Similarly, in Sweden once the Government started to pay teachers by results, student performance (as assessed by the teachers) began to steadily rise. The Government claimed this as a success of the policy, but others argued more convincingly it was grade inflation (Wikström, 2005) and that standards had not actually risen. Results from international surveys and other research support the argument that teachers were being incentivised to artificially boost results and not necessarily to improve learning. If (like the Dutch) the Swedish system had introduced external confirmatory assessments and validation of teachers' assessments, this uncertainty might have been avoided.

Sometimes these pressures can be great enough to incite criminal behaviour and there have been several ugly examples from the United States where teachers and public officials have been caught cheating - most recently the prosecutions in Philadelphia (2015) following wide-spread evidence of teachers and school principals involved in cheating on standardised assessments and the Atlanta Public Schools cheating scandal (2009).

One of the best studied is the Chicago Public School System cheating in the 90's (Jacob, 2003; Gay, 1990). In 1993, the Chicago Public School System introduced monitoring assessments, set externally but administered and marked by the teachers. In 1996, these assessments became high stakes tests when the authorities coupled these results to teacher performance evaluation and school funding. That year saw a spike in teacher-led cheating in these tests (it is customary to run statistical checks to highlight schools that make unlikely gains in performance). Analysis of the teachers detected cheating - only the ones uncovered by the algorithms and therefore not all the teachers who cheated - indicated they tended to be from poorer performing schools or were newly qualified teachers. These were teachers with the highest incentives to cheat. Alternatively, they may just have been the teachers most likely to be caught cheating due to their incompetence or inexperience.

The impact of this on the education system is a direct degradation - those schools are turning out students of unknown quality and the quality of learning in those schools has become irrelevant - that sets the conditions for loss of confidence and failure of the system, as eventually more schools are compelled to game the system and adjust their own results.

Rapid steps were taken in Boston to ensure the integrity of the assessments and following the introduction of punitive measures, including sacking teachers caught cheating, the incidence of detected cheating dropped by 30% - but the later cases in Atlanta and Philadelphia show that the pressures and incentives to cheat in standardised tests remain strong when the results from these tests are coupled to accountability measures or funding. This is a really important message for those engaged in education reform.

These issues are also true for primary examinations. SATs[5] in the UK for Students at age seven and 11 are meant to be a monitoring assessment, but the linking of the SATs results at Key Stage 2 (aged 11) to accountability measures has made them high stakes for schools. There is evidence

[5] SATs – Standard Assessment Tests – the common term for the National Curriculum tests in the UK to assess the attainment of children aged seven and 11-year-attending government maintained

that they have become *de facto* high stakes assessments and many teachers would argue that this has been very damaging for learning and many learners (and teachers) feel under pressure to perform well, with some schools adapting the curriculum prior to the assessments to just focus on what is being tested in the exam (Hutchings, 2015): “Year 6 pupils do no other subjects than literacy and maths from September until SATs.”

Similarly, the Secondary Entrance Assessment (SEA) in Trinidad and Tobago (Ministry of Education, 2013) - an end of primary test used to place a small minority of students into secondary schools, especially the denominational schools which have a very high demand from parents - is meant to be a voluntary examination aimed at differentiating the best students. It casts a long shadow over education in Trinidad and Tobago and the high-stakes, examination-focussed education it engenders competes directly against the Government’s reforms and attempts to improve the equity and the quality of education generally (De Lisle, 2012). Parents are under pressure to enter their children for the examination, and even schools whose student intake has little chance of achieving entrance to the elite secondary schools, feel under pressure to perform well in the examination as the school’s perceived quality (and hence their attractiveness to parents and the best students) is judged on the results. Schools adapt their teaching from early ages to performance on the test and there even exists pre-primary crammers whose selling point is early preparation to get students into the right primary schools, which have the best chance of propelling the student via the SEA into the best secondary schools. The impact of the presence of this high-stakes test in the system is predictably all the negative backwash effects one would expect: narrowing of the curriculum, rote learning, student stress, and disengagement from education. Sadly, because the examination has such a historical standing and support from influential sectors of society whose children benefit from the presence of the SEA, its abolition would require huge political confidence and strength of will by any incumbent education minister, and so this examination acts as a strong block to systemic change and the effective implementation of reform.

Assessment as both a driver and a block to change

All these case studies show that we need to be cautious when using assessment results as a proxy for education and that any systemic study or reform of an education system needs to include the existing examinations and assessments as key and powerful driving forces. System reform can only happen if the various assessments align with the reform. The outputs from these assessments also need to act correctly within the system feedback loops to push change in the correct way. If they do not, or the results are used simplistically to force change, then they are likely to become corrupted and, instead of promoting improvement, either derail or block change. In the presence of uncertain or contradictory information, systems are likely to revert to the status quo.

This is categorically not to say that examinations and assessment are an evil necessity in education systems. They are instead a very vital part of the education ecosystem providing vital information and feedback to allow autoregulation[6]. As the example of Flight 447 illustrates, without good information it is impossible to safely adjust or regulate a system. Without good measurement of learning, the system exists in a state of ignorance and government, schools and parents have to make decisions based on inadequate or misleading proxies - a child’s handwriting becomes more important that what they write, the quietness of a class more desired than whether they are learning, a certificate more valuable than ability.

Bad assessment not only tolerates or encourages poor teaching and learning, but poor assessment means that it is more likely that a student will be judged by their school or peer background in the absence of better evidence. The impact on the system can be disastrous, and ultimately the wrong people with the wrong skills end up in the wrong places. This reduces equity and wastes talent - as the economist George Stigler appositely puts it: “In a regime of ignorance Enrico Fermi would have been a gardener, Von Neumann a checkout clerk”.

[6] Autoregulation is a process in a biological system which allows an internal adaptive mechanism to adjust (or mitigate) that system in response to change in the system.

References

Adnan, U., Mahmood, M.A. (2014). Impact of Public Examination on Teaching of English: A Washback Perspective. *Journal of Education and Practice* 5(2).

Ahmed, Ayesha et al (2015). 1st ed. [pdf] AQA. Available at: http://filestore.aqa.org.uk/pdf/AQA-THE-FUTURE-OF-ASSESSMENT.PDF.

Black, P. and Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), pp5-31.

Campbell, Donald T. (1976). *Assessing the Impact of Planned Social Change.* The Public Affairs Center, Dartmouth College, Hanover, New Hampshire, USA.

De Lisle, J. (2012). Secondary School Entrance Examinations in the Caribbean: Legacy, Policy, and Evidence within an Era of Seamless Education. *Caribbean Curriculum,* Vol 19, pp109–143.

Gay, G.H. (1990). Standardised assessments: irregularities in administering of assessments affect assessment results. *Journal of Instructional Psychology*, 17(2) pp93-103.

Gibb, Nick (2011). Minister of State for Schools, DoE press release.

Goodhart, Charles (1984). Monetary Theory and Practice. 1st ed. London: Macmillan.

Hutchings, Merryn (2015). 1st ed. [pdf] Available at https://www.teachers.org.uk/files/exam-factories.pdf.

Jacob, B. and Levitt, S. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics*, 118(3), pp.843-877.

Kamrani, S. (2011). Future of Pakistan in respect of Education, Islamabad. Aziz Publishers

Ministry of Education, The Government of Trinidad and Tobago (2013). Available at http://moe.edu.tt/learning/primary/sea-exams.

Mirza, M., Nosheen M., and Masood N. (1999). Impact of Examination System on Teaching Styles of Teachers at Secondary and Higher Secondary Classes, Lahore. Institute of Education and Research, University of the Punjab.

Mundell, Robert (1989). Uses and Abuses of Gresham's Law in the History of Money, Columbia University.

Rehmani, A. (2003). Impact of Public Examination System on Teaching and Learning. *Pakistan International Biannual Newsletter ANTRIEP*, 8 (2) pp3-7.

Stacey, Glenys (2014). speech to Federation of Awarding Bodies. Available at https://www.gov.uk/government/speeches/putting-validity-at-the-heart-of-what-we-do

UNESCO (2016). 11th Education for All Global Monitoring Report. Available at http://en.unesco.org/gem-report/

Waldow F. (2009). What PISA Did and Did Not Do: Germany after the 'PISA-shock'. *European Educational Research Journal,* 8(3).

Wikström, C. and Wikström, M. (2005). Grade inflation and school competition: an empirical analysis based on the Swedish upper secondary schools. *Economics of Education Review,* Vol 24, pp309–322.

Wolf, A. (2011). 1st ed. (pdf). The Wolf Report. Available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/180504/DFE-00031-2011.pdf