RISE
RESEARCH ON IMPROVING SYSTEMS OF EDUCATION

WORKING PAPER
December 2017

Review of High Stakes Examination Instruments in Primary and Secondary School in Developing Countries

Newman Burdett

RISE-WP-17/018

Oxford Policy Management
BLAVATNIK SCHOOL OF GOVERNMENT
UNIVERSITY OF OXFORD
Center for Global Development

Funded by:
UKaid from the British people
Australian Aid

Review of High Stakes Examination Instruments in Primary and Secondary School in Developing Countries

Dr Newman Burdett

Newman Burdett
12/14/2017

Contents

Executive summary

The purpose of this investigation is to gain greater understanding of high-stakes examination instruments (i.e., tests used for progression or certification) in primary and secondary school in developing countries. This allows us to be better informed as to their potential as a lever for reform. The RISE working paper 16/010, *The Good, the Bad, and the Ugly - Testing as a Key Part of the Education Ecosystem,*[1] explored how assessment, especially in high stakes examinations, is an integral part of an education system and exerts a strong influence on what actually happens in the classroom. Others have also identified assessment as an important lever for reform.[2,3,4] However, for assessment to be an effective lever for improving learning outcomes, the assessment must be tightly coupled to (i.e., effectively measure) the desired learning outcomes. This paper looks at how effectively high stakes examinations in a small sample of representative countries (two in Africa and two in South Asia) perform in terms of assessing higher-order skills as well as reviewing, generally, how well they perform as assessment tools.

The paper starts by looking at some exemplar assessments (TIMSS, PISA, US, and UK high stakes examinations) to identify higher-order skills and construct a framework tool. It then investigates the levels of cognitive skills measured by a selection of high stakes examinations for primary, lower secondary, and upper secondary schooling in Uganda, Nigeria, India, and Pakistan. It also looks at the assessments used in Alberta, Canada, to provide a comparison from a developed world context. The four developing world countries were chosen as they have readily available assessment materials in English. Alberta was chosen as it performs well in international studies and has readily available assessment materials in English.

This paper studies how, and the degree to which, basic and higher-order thinking skills are required in examinations. These higher-order thinking skills go beyond the memorisation and recall of facts and are important as they are key skills in preparing students for the outside world, allowing them to adapt and use the knowledge they learn in school in more generalised contexts.

There is obviously an important caveat in this approach when applied to benchmarking and standards in that 'difficulty' is not a uni-dimensional construct and this paper only looks at one dimension in that space. For example, *'Describe the basic concepts of the theory of relativity'* is probably a 'difficult' question for most people, but it is a lower order skill in that it only requires recall. It is entirely possible to drill a student to answer this question without them having any understanding of relativity or even what some of the key phrases and words mean. While this paper focuses on a single dimension of examinations – the extent to

[1] http://www.riseprogramme.org/content/rise-working-paper-16010-good-bad-and-ugly-testing-key-part-education-ecosystem

[2] Braun, et al, Improving Education Through Assessment, Innovation, and Evaluation, 2006, https://www.amacad.org/publications/braun.pdf

[3] Kellaghan, Thomas; Greaney, Vincent. 2004. Assessing Student Learning in Africa. Washington, DC: World Bank. © World Bank. https://openknowledge.worldbank.org/handle/10986/14910

[4] Bethell, George. 2016. Mathematics Education in Sub-Saharan Africa: Status, Challenges, and Opportunities. World Bank, Washington,
DC. © World Bank. https://openknowledge.worldbank.org/handle/10986/25289

which items assess basic or higher-order skills – issues related to curriculum content and its alignment with examinations remain to be studied to create a more comprehensive picture.

The reason this study focuses on higher-order skills is to attempt to shine a light on how much ‘real’ learning is going on. Students may well be leaving examination halls having scored high marks, but not having learnt anything of use outside that examination hall. They could potentially be leaving school with no useful skills and poor literacy and mathematical skills beyond the very limited repertoire needed to pass the examination. However, if the examination requires them to not only recall knowledge, but understand, apply, and be able to use that knowledge in novel situations, then it is likely that what they learn in school will be useful beyond the examination.

Studying a range of international benchmarks drawn from assessments used by the Organisation for Economic Co-operation and Development (OECD), the International Association for the Evaluation of Educational Achievement (IEA), the US, and the UK, a tool was developed to classify items based on whether they were assessing:

- Level 1 Recall – having to recall a fact or piece of knowledge
- Level 2 Apply – understanding that knowledge and applying it
- Level 3 Reasoning – critically analysing and evaluating facts and potentially putting those pieces of knowledge together in novel ways

The content of each examination paper was then classified according to this tool. In cases where the level of the question was ambiguous, the marking rubrics were used as confirmation of the level.

Overall, the assessment materials showed a very low proportion of higher-order skills. In India and Pakistan, higher-order skills were almost entirely lacking and the focus was very much on recall of very specific rote-learnt knowledge (e.g., the price a character in one book paid for a carpet or the specific phrasing of a term in a text book).

In the two African countries, this rote-learning approach seemed less extreme, but there was still a very heavy focus on rote learning of facts (e.g., many science examinations were just recall of facts with very little attempt to probe understanding of those facts or how they fit into the wider scientific field of knowledge or scientific literacy in the everyday world).

There were some good examples where clear attempts had been made to test higher-order skills, for example the Nigerian National Common Entrance Examination (for entry into Federal schools and colleges) focuses on assessing more than just recall of knowledge. Similarly, the Ugandan Primary Leaving Examination assesses a wide range of skills including higher-order skills. Even so, the amount of higher-order skills assessed was less than would be expected in an international benchmark.

While the exam boards in Pakistan and India appear to produce low quality papers, there are private organisations in those geographies (the Aga Khan University Examination Board [AKU-EB] and Educational Initiatives [EI Pvt] are respectively examples) producing high-quality assessments which assess appropriate higher-order skills.

All four countries showed a variety of issues in the assessment materials and all showed examples of poor assessment technique. The Uganda PLE and Nigerian NCEE showed much better quality control and examples of good practice. Across all the examples, the quality was very variable with some of the Pakistan board papers riddled with mistakes and unanswerable questions. In the case of the Indian, Pakistan, and Nigerian Senior School Certificate, the quality of the papers was such that it is likely that the ability of these papers to predict which candidates are the best in that subject (as defined by their ability to go on to further study or work) is very low and in some cases more akin to a lottery. It is clear that all four countries would benefit from technical support in improving the quality of assessment materials.

In India and Pakistan, the focus on examinations that reward rote-learning must be a major barrier to any educational reform unless the assessments are changed drastically, although the ability of the system to adjust to such a change is undetermined.

In Nigeria, there clearly exists some good practice but, as in the case of India and Pakistan, the format and quality of the school leaving examinations provides a barrier to wider change. If Nigeria is serious about pursuing educational reform, then the assessment systems need to be reformed to align better with the aspirations of policy.

In Uganda, there is a stated desire to improve learning outcomes and ensure learners leave school with useful skills. They are clearly trying to adapt their assessments to encourage that but, especially at the higher levels, more work needs to be done and the format of the examinations, generally, allows poor teaching practices to continue. That said, some subjects have made the transition and have adapted to produce good assessments that encourage the learning of higher-order skills. Some examinations (e.g., entrepreneurship and a general paper), seem specifically designed to encourage good learning and the acquisition of higher-order skills.

Alberta, as the comparison jurisdiction, shows a consistently high proportion of higher-order skills including at Grade 9, which contrasts with the developing world examinations. The items are also produced to a generally higher quality and lack errors or ambiguities and are written in accessible language. The alignment between the assessments and the curriculum aims is also clearly articulated and demonstrated in the items.

This study did not investigate the curricula that sit beneath these assessments, but in many cases the focus is on learning a large amount of potentially useless facts. In the case of the Indian and Pakistan assessments, they are almost exclusively assessing a large body of facts and intimate familiarity with text books that is of very questionable use.

In science, in all the developing world countries studied, the focus is on assessing large amounts of facts related to science, but much less on the ability of students to think scientifically or apply that knowledge in a useful way.

Based on the assessment material, it would seem that many of the curricula need to be altered to focus less on facts and more on being able to usefully engage with the subject matter. This would require a lot of material to be removed from the curriculum to accommodate more teaching time focussed on ways of thinking. Based on many of the facts being assessed in the tests seen here (often trivial or irrelevant details), this loss of material would not impact the quality of learning and would likely improve it, but the curriculum would need to be analysed first. It could be that the intended curricula are decent, but the examinations are either misaligned and/or badly designed, drawing out only the useless aspects of an otherwise solid curriculum. It was not possible in this study to look at the curriculum that should be taught, but this would be an interesting further study as it would help understand how much poorly designed and implemented assessments act as a barrier to good education or whether they reflect poorly designed and implemented curricula.

Often what is being taught and assessed in the examinations studied for this paper is so divorced from what is expected to be taught and assessed in the international benchmarks, that making any meaningful comparison is impossible. It would be possible to score highly on one of the Pakistan papers, but fail to achieve the lowest benchmarks on IEA or PISA. Conversely, a good student from another system, for example the US or UK, could score highly in PISA or TIMSS, but fail to pass one of the CBSE or Pakistan Board Papers. The huge amount of very precise, extraneous recall required in these systems means that a significant part of the curriculum cannot be mapped onto any other benchmark system. The papers from CBSE and Pakistan do not match in any meaningful way what is considered necessary or good education in any of the available benchmarks.

The issues with many of the assessment instruments suggest that, if they are broadly representative of the developing world, there are issues with using national high stakes examinations to monitor educational outcomes.

Purpose and approach

The purpose of this investigation is to gain greater understanding of high stakes examination instruments in primary and secondary school in developing countries to be better informed as to their potential as a lever (or barrier) of reform.

This research set out to:

1. Develop a taxonomy of cognitive skills as a tool for categorizing examination items based on assessments such as U.S. Common Core State Standards, TIMSS, or PISA.
2. Using the above, within mathematics, English, and science subjects, detail the prevalence of cognitive skills examined.
3. Compare developing world assessments with examples from one developed world assessment (Alberta, Canada).
4. Investigate whether it is possible to conduct a rough mapping of an examination to a performance level on the anchor assessment.
5. Offer commentary and recommendations on how the examination instruments might be improved.

Countries

Test specifications from TIMSS, PIRLS, PISA, the UK, and the US were studied as examples of developed world assessments to define the skills and levels, and provide expected ranges for the various skills levels.

Uganda, Nigeria, India, and Pakistan were selected based on availability of assessment materials in English medium.

The latest released versions of items from Canada's Alberta Provincial Achievement Test at Grades 6 and 9 were also classified to provide a comparison with developed world jurisdiction high stakes examination papers. These were chosen because assessment materials are publicly readily available in the English language medium and Alberta performs well, but not exceptionally, in international surveys.

Why are higher-order skills important?

This paper sets out to better define what higher-order skills look like in the context of a developing world classroom and whether those skills are not only aspired to, but being effectively assessed.

Over the last century, education systems around the world have evolved and, especially in the last few decades, the number of children attending school has risen dramatically in the developing world. Unfortunately, this increase in enrolment has not necessarily led to improvements in the quality of education, governance, or pedagogy. Often curricula and assessments in developing countries lag far behind the developed world. This is exacerbated by changes in a labour market that increasingly requires a wider set of skills than basic literacy and numeracy.

The purpose of this paper is not to state the need for higher-order skills, as many others have already done that convincingly:

> *"For most of the last century, the widespread belief among policymakers was that you had to get the basics right in education before you could turn to broader skills. It's as though schools needed to be boring and dominated by rote learning before deeper, more invigorating learning could flourish.*
>
> *Those that hold on to this view should not be surprised if students lose interest or drop out of schools because they cannot relate what is going on in school to their real lives.*
>
> *….*
>
> *In 2010, the world is now more indifferent to tradition and past reputations of educational establishments. It is unforgiving to frailty and ignorant of custom or practice.*
>
> *We live in a fast-changing world, and producing more of the same knowledge and skills will not suffice to address the challenges of the future. A generation ago, teachers could expect that what they taught would last their students a lifetime. Today, because of rapid economic and social change, schools have to prepare students for jobs that have not yet been created, technologies that have not yet been invented and problems that we don't yet know will arise.*
>
> *Think back 50 years: could educators then have predicted how the Internet, which emerged globally in 1994, or the mobile phone, which appeared a few years later, would change the world? These technologies have not just become tools of learning, but networking and knowledge sharing, as well as innovation and entrepreneurship."*[5]

Rather, the aim of this paper is to develop a useful taxonomy of basic and higher-order skills and offer limited insight into the extent to which education systems in four developing countries emphasise these skills through their high stakes examinations. The importance placed on the results of these examinations – which determine whether students advance to the next stage of education or access employment – mean that they exert significant influence over both pedagogy and curriculum, and therefore over what students ultimately learn. By studying these examinations through the lens of basic and higher-order skills, we can gain useful insight into the type of learning education systems value and highlight areas for improvement.

Deciding on a taxonomy of skills

The first step in determining how well education systems in the developing world teach and assess higher-order cognitive skills, is to define more broadly the skills that are considered important and what differentiates desirable learning from less useful learning. To this purpose, this study initially looked to "21st Century Skills" (21C) as a good, though aspirational, starting point for developing a taxonomy of higher-order cognitive skills.

[5] http://www.oecd.org/general/thecasefor21st-centurylearning.htm

21C Skills are designed to represent a set of generically useful skills both in further study and the world of work. There is not uniform agreement on what 21st Century Skills entail but, for the purpose of this paper, we will assume the definition proposed by the project Assessment and Teaching of 21st Century Skills (ATC21S)[6], which maps well to definitions offered by other groups (Partnership for 21st Century Learning, Lisbon Council, ISTE NETS, etc.[7]). ATC21S, a group of 250 researchers from around the world, organised the skills into four broad categories: ways of thinking; ways of working; tools for working; and skills for living.

One of the issues with using 21C Skills for this study is that most of the curricula and assessment systems in our four countries will have been designed according to more traditional educational philosophies and theories. Much of the 21C Skills are quite sophisticated and require a skilled teaching force and complex assessment. These are often lacking in the developing world and pose challenges even in the more developed world. As Cambridge Assessment, a leading international assessment provider, admits when reviewing their own Thinking Skills Assessment: "writing appropriate item types that differentiate appropriately is an exceptionally difficult and skilled job, requiring significant training and a great deal of experience. This is not an assessment method for wide scale testing of 21st century skills across many students."

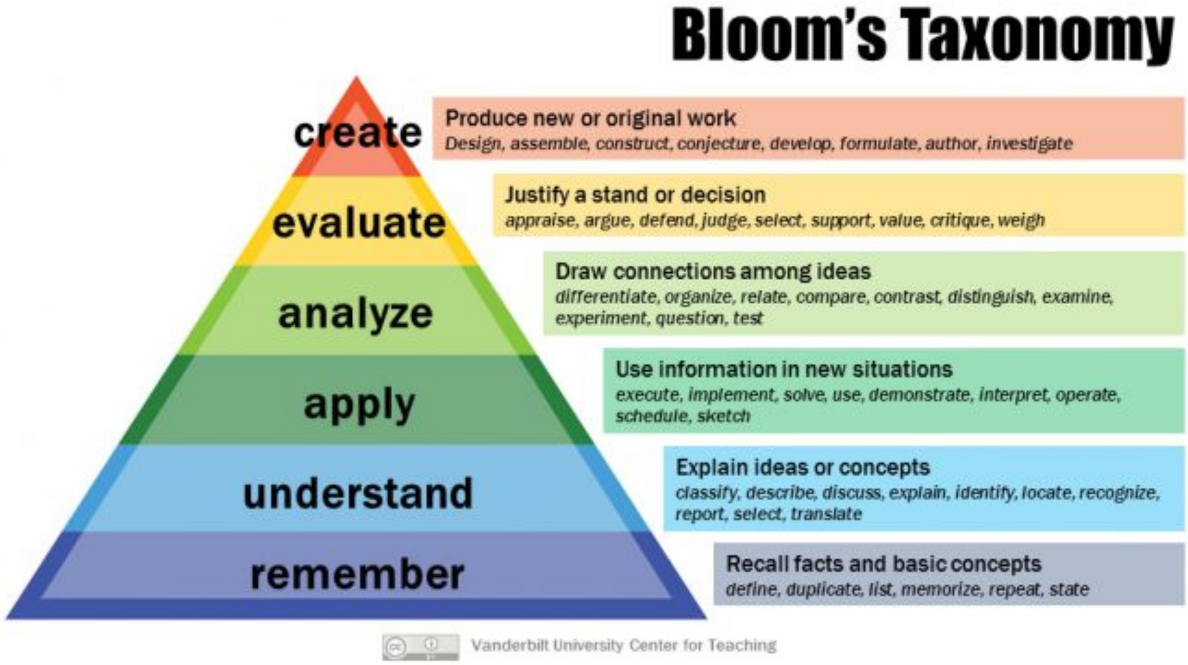
The second issue is how applicable these skills are for students in the developing world, a significant proportion of who have not achieved even basic literacy and numeracy, prerequisites to other higher-order forms of learning. In many of these countries, the average students are scoring well below the standard shown in the countries in which the 21C Skill were developed, so it is unreasonable to expect that good evidence of higher level 21C Skills will be found in the education systems of these poorly performing countries. To identify skills appropriate for the developing world we need to ensure that our criteria are suitable for the context.

Fortunately, Bloom's Revised Taxonomy[8] provides a more appropriate foundation for this study and is frequently used in curricula and assessment design. It links well to the 21C Skills and details the foundational skills needed for many of 21C Skills.

[6] Griffi n, P., & Care, E. (2015). The ATC21S method. In P. Griffi n & E. Care (Eds.), Assessment and teaching of 21st century skills: Methods and approach

[7] The Cambridge approach to 21st Century Skills: definitions, development and dilemmas for assessment. Suto, I. and Eccles, H. IAEA 2014 Singapore

[8] A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives: editors, Anderson, L. W., Krathwohl, D. Longman 2011 NY


Bloom's Taxonomy
create
Produce new or original work
Design, assemble, construct, conjecture, develop, formulate, author, investigate
evaluate
Justify a stand or decision
appraise, argue, defend, judge, select, support, value, critique, weigh
analyze
Draw connections among ideas
differentiate, organize, relate, compare, contrast, distinguish, examine, experiment, question, test
apply
Use information in new situations
execute, implement, solve, use, demonstrate, interpret, operate, schedule, sketch
understand
Explain ideas or concepts
classify, describe, discuss, explain, identify, locate, recognize, report, select, translate
remember
Recall facts and basic concepts
define, duplicate, list, memorize, repeat, state
Vanderbilt University Center for Teaching

In the revised taxonomy, knowledge is at the basis of all cognitive processes, which form a hierarchy with the higher-order skills at the top of the pyramid.

International framework benchmarks

The other advantage of using Bloom's taxonomy as the basis for developing the framework tool, is that Bloom maps well to many of the international benchmarks. Several systems were studied for this paper including the Common Core State Standards in the USA, the UK frameworks, and the PISA frameworks. These systems all emphasise higher-order skills and agree broadly on what those skills are, although they differ in the details and balance.

The UK and USA frameworks are hard to apply as they provide a high-level framework within which various assessment systems work and it is these sub-systems that then define the frameworks by interpreting them in a variety of ways. In the UK, it is individual awarding bodies that translate the frameworks into assessment specifications and in the USA, it is individual states and assessment bodies (e.g., the Smarter Balanced Assessment Consortium and the Partnership for Assessment of Readiness for College and Careers [PARCC]). In some cases, the frameworks are constructed as curriculum standards linked to curriculum content and this, coupled with the variation in interpretations about how the skills are defined, make the UK and USA frameworks less applicable as more generic frameworks.

The PISA frameworks are designed to be independent of the curriculum they are testing and so would seem ideal as an international benchmark, but PISA uses a relatively sophisticated test specification that employs several dimensions to define an item. This makes it harder to use when applied to non-PISA items and potentially less reliable, and therefore not well suited as a benchmark for this study. OECD are also developing a version of PISA for the developing world, PISA-D, and once technical details about how the frameworks will be applied are available, it might be worth reconsideration. All current indications are that PISA-

D will be built a similar way to PISA and would be hard to apply reliably *post hoc* to other systems.

The IEA frameworks (such as TIMSS and PIRLS) have the advantage of being simple and robust to employ. They also fit well with the various criteria discussed above and across subjects. Similarly, the proposed framework (see below) fits well with the IEA definitions in science, mathematics, and reading literacy. The IEA frameworks also are intended to work in the developing world and PIRLS Literacy and the Literacy and Numeracy Assessment (LaNA) will be linked to the TIMSS and TIMSS Numeracy assessments conducted in 2015 and to the PIRLS and PIRLS Literacy assessments conducted in 2016. The exact details of the frameworks are not yet released, but they state that, like with the OECD PISA-D, they will use the same specification framework as the main studies.

Study of the various frameworks show that whilst there are considerable differences in the details of how each organisation defines and categorises the cognitive domains, there is broad agreement on what constitutes low, medium, and high levels of cognitive demand – at least as far as drawing up a framework to analyse national assessment systems – that fits reasonably well with the framework proposed here: recall, understanding, and high-order skills.

The table below shows how the various frameworks classify the levels and how they fit against the levels proposed.

Skill order	IEA Maths[9]	IEA Science	IEA PIRLS/PIRLS Lit.	PISA Lit.[10]	PISA Sci.[11]	PISA Maths
1	Knowing	Knowing	Focus on and retrieve explicitly stated information	Locate information	Low - Carrying out a one-step procedure, e.g. recall of a fact or locating a single point of information	Formulating situations mathematically[12]
2	Applying	Applying	Make Straightforward Inferences	Understand	Medium - Use and application of conceptual knowledge to describe or explain phenomena, select appropriate procedures involving two or more steps, organise/display data, interpret or use simple data sets or graphs.	Employing mathematical concepts, facts, procedures
3	Reasoning	Reasoning	Interpret and integrate ideas and information Evaluate and critique content and textual elements	Evaluate and reflect	High - Analyse complex information or data, synthesise or evaluate evidence, justify, reason given various sources, develop a plan or sequence of steps to approach a problem.	Interpreting, applying and evaluating mathematical outcomes

[9] 2011 framework
[10] 2018 framework
[11] 2015 framework
[12] As PISA uses a more complex matrix to define skills there is potential overlap in how these criteria are defined and so it falls into both categories and definition may depend on context of individual items

Balance of skills

Alongside the variations in how the frameworks categorise the skills, there are also variations in how the frameworks think the skills should be balanced. This is demonstrated in the table below which shows the proposed percentage of test items at each level. The Oklahoma Core Curriculum Tests were chosen to represent the CCSS as they are easily available and the test specification is easily comparable to the proposed framework. These percentages should be read with the caveat that the actual demand of questions is more complex than this uni-dimensional simplification, but they do give a good indication of what can be expected in a high-quality assessment framework.

The PISA Science specification could not be adequately equated to the proposed framework and so is not included.

Assessment	Recall	Understanding	Higher-order
PIRLS	20	30	50
PIRLS Literacy[13]	50	25	25
TIMSS Grade 4	40	40	20
TIMSS Grade 8	35	40	25
PISA Math	25	50	25
PISA Literacy	25	50	25
CCSS (Oklahoma) Science	10-15	60-65	25
CCSS (Oklahoma) Math Grades 3-5	20-25	65-70	10
CCSS (Oklahoma) Math Grades 6-8	10-15	65-70	20

The PIRLS Literacy assessment is aimed at the lower end of the achievement scale and so probably best reflects the level of learning in the contexts this framework is designed to study.

Overall, the IEA frameworks seem to require more recall based on these figures and this is another reason for using them as a benchmark for developing a framework tool as they are probably a fairer reflection of what can be expected.

The outcomes of this suggest that to meet international benchmarks, national assessments should have 10-20 percent of items aimed at assessing higher-order skills at lower primary level and, 20-25 percent at grades higher than this.

Developing and applying the framework tool

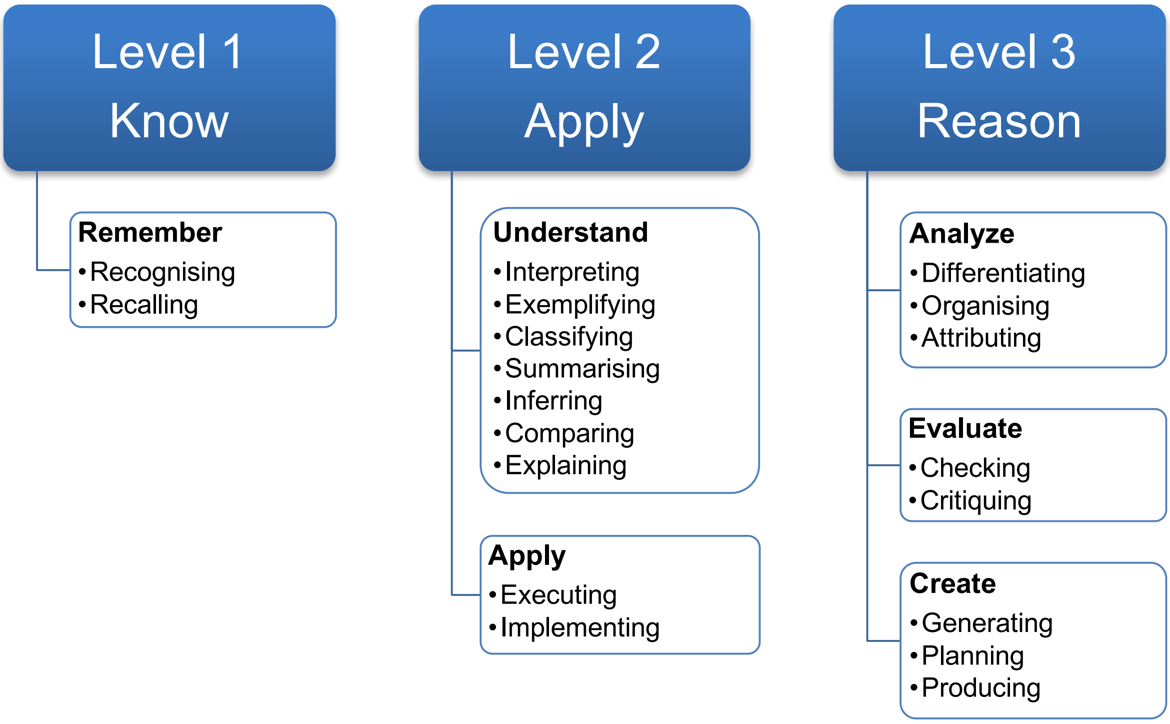
TIMSS and PIRLS frameworks were selected as the benchmark in developing the tool because they were considered the simplest and most reliable to apply in the context of the developing world and will allow mapping directly against LaNA at a later stage. To enable easier interpretation, the levels for PIRLS and TIMSS have been combined based on Bloom's Revised Taxonomy to give a single set of levels across all subjects. This means that

[13] Formerly known as pre-PIRLS its purpose is to extend the effective measurement of reading literacy at the lower end of the achievement scale at Grade 4

the framework is designed to apply across all subjects rather than have specialised criteria for each subject.

The following overarching framework provides a tool for evaluating the competencies being assessed and is mapped directly onto the IEA frameworks using the following definitions. The bulleted points are key-words for guidance and exemplification and not intended as an exhaustive list.

It is important to note that in terms of the learning process, this framework is not static and as learners progress some skills that were higher-order (e.g., differentiating numbers into odd or even), would be expected to become straight recall. This means that care needs to be taken in applying this framework tool to consider the context and age of the learners.



Level 1 Know

Remember
- Recognising
- Recalling

Level 2 Apply

Understand
- Interpreting
- Exemplifying
- Classifying
- Summarising
- Inferring
- Comparing
- Explaining

Apply
- Executing
- Implementing

Level 3 Reason

Analyze
- Differentiating
- Organising
- Attributing

Evaluate
- Checking
- Critiquing

Create
- Generating
- Planning
- Producing

In most cases, the categorisation will be based on the command word (the word or phrase that instructs the candidate what to do). For example, if the question is: ‘*State the formula for water*’, this would be a clear Level 1 (recalling). Similarly, ‘*Read the above paragraph and produce a short (1 paragraph) summary of the main points.*’ would be Level 2 (summarising).

Care needs to be taken in applying the taxonomy in that the correct command words are not always used. For example:

> Explain the terms fats and minerals [5 marks]

This is actually asking candidates to define the two terms ‘*fats*’ and ‘*minerals*’, a recall task and, hence, Level 1. There is no explanation required for this answer.

This misuse of command words is a common problem with many examinations. For example, ‘*explain*’, is often incorrectly used when the question and marking scheme are actually testing recall of facts, rather than requiring the learner to provide a genuine explanation.

If the question and marking scheme rubric are misaligned, then this is potentially a problem for the validity of the test items and raises questions about the quality of the assessment. If the test specification requires a certain proportion of Level 2 or 3 items, then a misaligned mark scheme might reduce the level of demand. The answer on the mark scheme might not align at all with what the question is asking, making it very difficult to interpret the learner's score.

Where there is doubt, the Level will be defined by whether the candidate needs to simply recall facts to answer the question (as above) or needs to interpret or do something more to get to the answer.

Differentiating Level 3 from Level 2 is sometimes more ambiguous. Where the task is ambiguous in terms of demand according to this taxonomy, and it is not clear whether it meets the criteria for Level 3, it will be cross-checked against the marking scheme (a rubric to markers which instructs them which answers are to be accepted) to ensure that the task requires the student to engage in relational thinking (i.e., the student has related, linked, or integrated the ideas) or for the student to use extended abstract reasoning (i.e., needs to take related ideas and extend them). [14]

For example, if a learner is required to write a letter applying for a job, the level of cognitive demand is determined not by the task, but by how the quality of the letter is assessed. If the answer is assessed on the creativity of the content, how well it addresses the issues, using a variety of sources, and requiring some form of organisation of ideas and evaluation of positions, then this would count as a Level 3 answer. However, if a letter is marked according to the scheme below it would be a mix of Level 1 and Level 2 marks.

[14] Biggs, J. and Collis, K. Evaluating the Quality of Learning: The SOLO Taxonomy New York: Academic Press, 1982

INVITATION LETTER TO INVITE LOCAL DIGNITARY TO SCHOOL PRIZE GIVING

You are the Secretary for the School. Write a letter to the Director of Education inviting him to your School Prize Giving, to take place at 3pm Saturday 13 August in Marshal Secondary School, Canton Road. Request his presence to present the prizes and give a small speech. Request confirmation of his attendance.

MARKS-5

Objective: To use the given input in a short, sustained piece of writing

Marking:

Content – 3 marks – (the given information in the question paper to be included)

– event - details

– purpose of invite - request to grace occasion

– date / time / venue

– request confirmation

Format - includes date, subject, addressee and closing. Format to be treated as part of the content

Expression - 2 marks (fluency and accuracy to be taken into account)

For the content, Level 1 marks are for the format as it is direct recall of knowledge on how to format a letter, where to put the date, signature, etc. The rest of the content is for ensuring the information in the question is included, which is potentially Level 2 as they need to recognise and extract the right information. In this example, the information is very clearly laid out and explicit, and therefore is low demand, possibly Level 1. At no point is the learner really required to engage with the subject matter on a deeper level and can gain full marks for repeating the information given in the question. The expression marks refer to spelling, punctuation, and accuracy of English, and so are Level 1 and 2 marks.

Similarly, for mathematics ‘*prove*’ could be a Level 3 skill as it should require learners to demonstrate relational thinking or extended abstract reasoning. Mostly, however, ‘*prove*’ is used to mean ‘*show*’, an application Level 2 skill. For example:

Prove that Sin 60^{o} = $2\text{Sin}30^{o}$ x $\text{Cos}30^{o}$ [5 Marks]

This question only requires the recall of various values and simple calculation. In this case the mark scheme is:

Sin30^o = ½ [1]; Cos30^o = √3/2 [1]; 2x1/2x√3 /2 = √3/2 [1]; Sin60^o= √3/2 [1];

therefore, Sin 60^o = 2Sin30^o.Cos30^o [1].

In fact, learners are very likely to have already encountered and learnt this specific relationship (or one very similar) which further would reduce the skill level, possibly making it a Level 1 skill if they have rote learnt it.

Generally, for Level 3 to be assigned, the question must be something that genuinely requires the learner to use higher cognitive skills to answer it, for example a novel context, drawing conclusions from new information, or planning and organising ideas in different ways. At the lower grades, these do not necessarily need to be profound, but they do require a certain amount of genuine analysis or creativity.

In this study, with only one evaluator, any ambiguities in applying the framework do not lead to inter-rater variation. If it were to be applied to other studies with several evaluators, obviously efforts would need to be made to align raters or use multiple raters for each item to ensure reliability.

Overview of examinations materials by country

Unless stated otherwise, the 2016 papers for English, mathematics and science(s) were analysed according to the above framework tool. Each marking point was assigned a Level from 1 to 3 based on the framework tool and an overall percentage of these levels was determined for each subject.

Uganda

In Uganda, the high stakes examinations are produced by Uganda National Examinations Board (UNEB) under the Ministry of Education & Sports. This review looked at three levels of examinations corresponding to Grades 6, 10, and 12; the Primary Leaving Examination (PLE), the Uganda Certificate of Education (UCE), and the Uganda Advanced Certificate of Education (UACE), respectively.

At Advanced Certificate Level, the General Paper was taken as a proxy for English as it assesses candidates' comprehension and communication skills.

The overall percentages for the three examinations are shown in the tables below.

Primary Leaving Examination (% by Skill Level)			
	English	Mathematics	Science
Level 1 Know	0	8	77
Level 2 Apply	75	80	23
Level 3 Reason	25	12	0

Uganda Certificate of Education (% by Skill Level)						
	English	Mathematics	Physics	Chemistry	Biology	Science (avg.)
Level 1 Know	0	0	41	46	44	44
Level 2 Apply	80	97	59	49	48	52
Level 3 Reason	20	3	0	5	8	4

Uganda Advanced Certificate of Education (% by Skill Level)						
	GP	Mathematics	Physics	Chemistry	Biology	Science (avg.)
Level 1 Know	21	0	34	32	35	34
Level 2 Apply	68	100	66	66	55	62
Level 3 Reason	11	0	0	2	10	4

Some of the primary papers are closest to what we would expect according to good international practice with a balance between knowledge, recall, and reasoning. The mathematics and English papers are close to the expected range (10-20%), although in science, the items are mostly testing lower order skills and a lot of direct recall.

At the Certificate of Education and Advanced Certificate level, only the English test papers have a significant proportion of higher-order skills close to the range expected from the various developed world and international assessments studied to produce the framework tool (20-25%). At these higher levels, the other subjects have low proportions of higher-order skills.

It is not clear if this is because of the higher stakes nature, an overcrowded curriculum causing an over-emphasis on recall, or whether the current round of reforms that the Ugandan government is pursuing has yet to filter through to these higher levels.

Where the English items are testing higher-order skills they are, for the most part, well-structured with the comprehension items requiring the candidates to actively engage with the text and synthesise knowledge from various paragraphs or summarise the ideas in the text. For example, one of the Primary level examinations in English requires learners to study a table of data about class borrowing habits compiled by a fictitious librarian and answer questions on it ranging from extracting information from the table (e.g., '*Who is the librarian?'* or *'Who borrowed most books?'*) to using the data to draw conclusions (e.g., *'Who is most likely to buy a new book?'* which requires the learner to note that one child has lost a book and will need to replace it).

The Advanced level includes a General Paper, which was assessed as there is no compulsory English paper (and it assesses many of the same skills). This paper includes some questions designed to assess higher-order skills such as analysis, planning, and evaluation. Unfortunately, this is a question paper where the learner can choose between various questions and the different options have different balances of skills. In the example studied, one option was to analyse given data about various transport providers and bus routes and make a series of evaluations based on that data. This option had a high proportion of higher-order skills. The other option was a more straightforward, previously unseen English passage comprehension exercise and had a lower proportion of higher level skills.

In the sciences, higher-order skills are assessed in the 'practical' papers (papers designed to test learners' experimental skills) and one of the reasons that physics fails to assess these higher-order skills is that the questions are too highly structured with the candidate being told what to do, how to record data, and how to analyse it at each step. This meant that learners had no opportunity to demonstrate independent thought or decision making.

There are numerous examples within these papers where the command words used could put the level higher, but what the question is actually asking from the candidate is a lower level, for example many of the '*explain*' or '*discuss*' questions are really asking for the candidate to '*describe*', requiring only straight recall.

Similarly, the level required for some of the questions can depend on the candidate's answer (e.g., a PLE English where they are asked to suggest a title for a poem about football that could be answered by the trivial 'Football', rather than something requiring more creativity).

The other difficulty in assigning levels to these papers is that many of the examinations have lots of alternative questions and routes through the examination. In the physics Advanced Certificate Paper 3 there are 324 different permutations of questions that a student could take with some of those options having over twice as much Level 2 application as other routes. This reduces the reliability and validity of these test papers.

The final issue in assigning levels to these question papers is that the format often makes it very easy for candidates to resort to rote learning material to answer the higher-order questions. For example, the format of the English open response questions allows for candidates to learn generic essays or letters that will still score highly. The essay titles are topics such as:

- Write a story that starts, 'The day finally arrived…'
- Money does not make you happy. Do you agree or disagree? Give reasons.
- Narrate an incident that made you a hero.

The openness and predictability of these titles coupled to the large amount of choice (one from six titles) does mean that a strategy of learning generic essays and practiced responses is likely to allow candidates to score highly. Similarly, the mark schemes for the letters are very generic and students are more likely to score well by being taught examination technique (i.e., how to construct a letter for the exam and to include the information given in the question and required in the answer), than through any creativity or real application of higher-order skills.

The report on the 2016 examinations from the UNEB website highlights concerns that too many students seem to be reproducing learnt responses and that teachers need to change the way they approach teaching these skills.[15] As discussed above, the format of the examinations does not discourage teachers using rote instruction and teachers might be disadvantaging their children by taking a less examination oriented approach.

[15] http://uneb.ac.ug/downloads/2016_UCE_RELEASE_STATEMENT.pdf

The overall quality of the papers in Uganda, in terms of production, was among the best seen in all four of the developing countries studied (i.e., no obvious errors and clear, unambiguous language). However, whilst the UNEB examinations were amongst the highest quality of those reviewed, it is likely that the numerous routes through many of the papers undermine the ability of the examinations to reliably assess candidates based on higher-order skills and may discourage teaching of those skills.

That said, it is clear from the materials studied that there is an attempt to provide a relevant education and produce assessments that support this. For example, Uganda also has entrepreneurship as a school subject. This is a mix of knowledge around business procedures and terminology, but also requires students to engage in entrepreneurial activities (a school Business Club project) and a work attachment or field trip. The examination assesses learners on their evaluation (a higher-order skill) of these and on a given case-study, as well as testing their business knowledge.

Nigeria

Examinations in Nigeria are conducted by the West African Examinations Council and the National Examinations Council. The main school leaving examination is the Senior School Certificate Examination (SSCE) at Grade 12, although its predecessor, the General Certificate of Education (GCE), is still offered as a fall-back to students who failed to get the required credit in the SSCE to progress. There is also a Basic Education Certificate Examination (BECE) at the end of Junior Secondary (Grade 9) and a National Common Entrance Examination (NCEE) at the end of Primary (Grade 6) for entry into the Federal Secondary Schools and Colleges - which are meant to be better than the state secondary schools and have more highly-qualified teachers. Science SSCE Grade 12 papers were not available for study.

The overall percentages for the three examinations are shown in the tables below.

National Common Entrance Examination Grade 6 (% by Skill Level)			
	English	Mathematics	Science
Level 1 Know	31	6	100
Level 2 Apply	48	78	0
Level 3 Reason	21	16	0

JSS BECE Examination Grade 9 (% by Skill Level)			
	English	Mathematics	Science
Level 1 Know	65	7	87
Level 2 Apply	34	93	13
Level 3 Reason	1	0	0

SSCE Grade 12 Examination (% by Skill Level)			
	English	Mathematics	Science
Level 1 Know	34	4	na
Level 2 Apply	66	95	na
Level 3 Reason	0	1	na

The National Common Entrance Exam is of noticeably higher quality than both the BECE and SSCE according to the proportion of higher-order skills being tested, as well as in terms of quality control, language used, etc. The science NCEE seems to test only recall of facts, but the English and maths papers have a good range of questions.

The NCEE assesses quantitative aptitude in addition to mathematical skills. There is a flaw in the way that this is structured in that many of the questions involve the student having to analyse and decode a particular problem and this then repeats twice more with the same logic but different numbers.

For example:

> Study the sample and work out the rule:
>
> 1 1 2, 6 36 42, 4 16 20
>
> What is the answer to:
>
> 1. ? 25 30 A. 55 B.45 C.25 D. 15 ***E. 5***
>
> 2. 9 ? 90 A. 27 B.36 C.45 D.64 ***E. 81***
>
> 3. 3 9 ? A. 6 ***B.12*** C.18 D.27 E.36

The first part, analysing the sample numbers and working out the rule in the sequence (the first number is squared to get the second and then the first and second added to get the third), is a Level 3 skill, but the second two questions would seem slightly redundant in that they just require application of the same reasoning (or if the learner failed to decode the initial example then they would fail to get all three marks). But, by requiring learners to think about the numbers given and create and assess alternative rules to arrive at the correct answer, it still demonstrates one of the best attempts seen in the papers studied to effectively assess higher-order mathematics skills in any of the developing world systems studied.

Many questions, at all levels, would on the surface appear to be asking more than is really required and use words like '*explain*', '*discuss*' or even '*analyse*' when they are really asking for a recall of facts and would be better phrased as '*state what you know about …*'.

Other than this, the examinations are generally well prepared but, especially in the BECE and SSCE papers, there are many instances of poorly phrased questions or mark schemes that might confuse the brighter candidates or at the least would randomly reward candidates as they try and guess the answer required.

This is illustrated in the examples given below taken from the English SSCE (the mark scheme answer is given in bold).

Most people abhor nagging. (a)scorn ***(b) despise*** (c) reject (d) slight.

Scorn, despise, and reject can all be synonyms of abhor, so how is a learner expected to choose the right answer?

He did all he could to fortify the building against attack. (a)defend (b)secure ***(c)shield*** (d)support

Again, *a, b,* or *c* could be correct, and *b* is possibly the best as *a* and *c* could alter the meaning of the sentence subtly.

In one comprehension task, where candidates must explain why a father refuses to answer the door, the required answer is one of the weakest and arguably wrong. Better students would give other answers that more logically fit the question based on the whole context of the passage, rather than the flawed logic of the answer that relies on the conversation that follows the father's refusal to answer the door.

These badly written questions mean that better candidates, those who understand the question and can apply higher-order skills, might be confused or give a correct answer deemed incorrect by the mark scheme, and hence penalised.

The problems illustrated above in the SSCE, were also found in the BECE where many items were so poorly written as to be unanswerable. For example, in science:

Which of the following is not found in aquatic animals?

(a) cilia

(b) flagella

(c) fur

(d) gills

(e) webbed feet

The presumably correct answer is *(c) fur*[16], but many aquatic mammals do have fur, such as fur seals, sea otters, etc., and conversely, many non-aquatic animals have characteristics that fall into other categories (e.g., humans have cells with cilia and flagella). This unintentionally disadvantages the brighter and more knowledgeable students who may try to deduce the most likely intended correct answer from a set of incorrect options. For the lower ability candidates, it is likely they will see 'fur' and choose that answer without further

[16] Unfortunately, the answer key was not available for this paper.

thought. In some cases, it was also not clear that the questions were testing the subject matter, for example in the English papers it was not uncommon for questions to be testing general knowledge rather than English vocabulary.

> Current law on human trafficking was (a) annulled (b) composed (c) enacted (d) imposed (e) created by the 8th National Assembly?
>
> Nigeria is rich in solid (a)bitumen (b)block (c)minerals (d)mines (e) particles

Without that specific background knowledge, the correct answers are not clear, even to a strong English speaker.

The lack of higher-order skills in the BECE and SSCE examinations, how the questions are structured, and the way they assess the candidates mean that they are unlikely to promote teaching of higher-order skills or of effectively distinguishing between candidates in any meaningful predictive manner.

The NCEE seems to be better produced and to higher quality standards, possibly because it is designed to select those students with most potential to go into the federal schools and colleges.

India

There are many examination bodies in India, including the various state examination boards and the Council of Indian School Certificate Examination, but the largest is the Central Board of Secondary Education (CBSE) which offers examinations at Grades 10 and 12. This study only looked at papers from CBSE as it is both the largest provider of examinations and had the most easily accessible assessment material.

The overall percentages for the examinations are shown in the tables below.

CBSE Grade X Examination (% by Skill Level)			
	English	Mathematics	Science
Level 1 Know	30	0	53
Level 2 Apply	70	100	45
Level 3 Reason	0	0	2

CBSE Grade XII Examination (% by Skill Level)						
	English	Mathematics	Physics	Chemistry	Biology	Science (avg.)
Level 1 Know	44	0	43	63	95	67
Level 2 Apply	55	100	57	37	5	33
Level 3 Reason	0	0	0	0	0	0

These papers are mostly aimed at testing recall and rote learning with very little evidence of any higher-order skills. The Grade 12 papers have no evidence of higher-order skills and the amount of direct recall is very high – indeed the science papers are mostly direct recall. In the mathematics paper, it could be argued that if learners are instructed using past papers that most of the questions classified as application could be recall.

Even in the literature and comprehension sections of the English papers, where there should be ample opportunities to test learners' higher-order skills, it is rote learning that is being tested. For example, the question below is taken from a CBSE English paper. The expected answers are given below the question.

> Read the extract given below and answer the questions that follow:
>
> So I went home and sat down before my desk and … waited, but nothing happened. Pretty soon my mind began to wander off on to other things.
>
> (a) Who is 'I' ?
>
> (b) What did 'I' wait for ?
>
> (c) What does the phrase, 'wander off ' mean ?
>
> Answers :
>
> *(a) John Hallock*
>
> *(b) inspiration to write a ghost story.*
>
> *(c) deviate / stray / be lost / get distracted*

It is clear from this that learners are meant to memorise and recall sections from literary works and are not required to engage with these texts in any more meaningful way. There are many examples like this and it is clear from the questions that the examinations expect only direct recall of facts from the set books.

> Why did 'Rev. McLeery' bring a rubber ring with him to the prison?
>
> *Answer: as part of escape plan for Evans*

The passage comprehension sections, which give an unseen passage in the examination paper, also score low on the skills levels as they mostly require direct matching rather than any comprehension of the text. For example, the question:

> How can computers help people going on holiday?

only requires the learner to identify the following sentence in the text and copy directly from it with no interpretation or synthesis required.

> "…For instance, *people going on holiday* could be informed about weather conditions."

In many of the comprehension questions the candidates need not even understand the question or relevant text, but by direct matching can still gain full marks. For example, many highly-educated, first-language English speakers would struggle to understand the following question:

> What prejudice has vitiated the reasoning of geologists?

But in the CBSE examination this can be answered by simply matching the words and lifting the relevant section from the text without any comprehension of its meaning whatsoever.

Even the longer literature analysis essay questions require direct recall of facts and memorising quotes from the book, rather than any analysis or evaluation of the text.

> Attempt a character sketch of Squire Cass.
>
> Answer:
>
> - *tall stout man of sixty, frowning face untidily dressed*
> - *signs of habitual neglect about him*
> - *self-possession and authoritativeness of voice and carriage*
> - *speaks in ponderous coughing fashion, lives an idle life*
> - *richest man in Raveloe*
> - *lazy, complacent, selfish and short tempered*
> - *father of Godfrey and Dunstan*
> - *seems to care more for his money than his sons*
> - *allows them to do whatever they please as long as they do not involve his tenants in any way*
> - *kept his sons at home in idleness – fell short in their upbringing long after wife's death*
> - *he condescended to preside in the parlour of the Rainbow - shows his arrogance.*
> - *fed dogs beef while commoners lined up for ale*
> - *sharp tongue*
> - *banishes Dunstan*

Most of the comprehension questions are poorly designed and encourage direct matching and lifting of words or phrases from the text. Often these are harder to answer if candidates use higher-order skills as they then need to work through several plausible options and reject them before answering, which involves more processing. For example, in one comprehension passage there is the line *"…cubs hidden away among rocks, hollows of trees, and other impossible places."*

> (a) To protect its cubs the mother panther hides them:
>
> (i) among rocks
>
> (ii) in the branches of the trees
>
> (iii) behind the tree trunks
>
> (iv) at its heels

The expected answer is (i) among rocks, but the context does not preclude options (ii) and (iii) and a candidate who reads and tries to understand the passage needs to do more mental work than one who just matches phrases. This is not good practice and actively encourages candidates to not engage with the text at a deeper level.

The composition questions requiring the candidate to write a letter and an essay are very generic, similar to the examples given for previous countries, and the letter writing exercises give marks for the procedural aspects of writing letters and for including named content, but there is no differentiation on quality of response beyond that. As with the previous examples, they encourage the learning of prepared responses and examination technique over useful learning.

In the mathematics papers, the assessment is all about application of mathematical processes. The questions are very formulaic and structured such that if the candidate has been taught how to approach the question, then they can answer it. This leaves no scope for deeper thinking or any creativity.

In science, attempts are made to assess higher-order skills, but on the whole, these fail. For example, the nominally open-ended question:

> Narrowly utilitarian arguments are put forth in support of biodiversity conservation. Explain the other two arguments that are put forth in support of the same cause.

The question assumes that there are only a limited and predetermined set of arguments and is clearly looking for a learnt response. Similarly, the questions aimed at assessing attitudes to science rely on candidates giving a very narrowly prescribed set of responses.

> You have a friend whose parents are too indulgent in his/her daily affairs. They think him/her to be still young which makes him/her sad and is upset all the time. As he/she feels that the parents should give him/her opportunity to take independent decision on some issues.
>
> (a) Would you support your friend and why?
>
> *Ans. Yes, because of peer understanding*
>
> (b) Write the characteristics of this age group.
>
> *Ans. curious, adventurous, looking for excitement, experimentation*
>
> (c) List two curative measures.
>
> *Ans. Avoid undue peer pressure / education & counselling / help from parents & peers / identifying the danger signs / professional and medical help or any other appropriate measures (any two)*

Setting aside that there is no science or scientific thought involved in this question, there is no acceptance that some students might answer ‘no’ and give valid, well thought-through reasons. Similarly, the characteristics are presumptive and not particularly linked to the context, and the curative measures are so broad as to include almost any possible answer.

The CBSE website also has some specimen papers to illustrate to schools what is required in the 2017 examination and, presumably, allow the schools to adapt to changes in content and format, although it was not clear what these changes are. The specimen 2017 papers are better structured than the 2016 past papers and, for example, the specimen 2017 chemistry paper (the only one with a setting grid included) identifies 10 percent of the marks as Higher-order Thinking Skills (HOTS). Analysis of the question identified as a HOTS question indicates that it requires significant thought to answer.

> A ketone A which undergoes haloform reaction gives compound B on reduction. B on heating with sulphuric acid gives compound C, which forms mono-ozonide D. The compound D on hydrolysis in presence of zinc dust gives only acetaldehyde. Write the structures and IUPAC names of A, B and C. Write down the reactions involved.
>
> (ii) Predict the products formed when cyclohexanecarbaldehyde reacts with following reagents.
>
> (a) PhMgBr and then H_3O^+.
>
> (b) Tollens’ reagent.

But this is part of an either/or style question and the alternative question is only a completion of missing reagents or conditions, and is mostly straight recall or limited application (albeit of some high-level content).

The English comprehension exercise in the specimen papers is also better than that found in the 'live' papers, including requiring candidates to identify information spread over different sentences or paragraphs, but still involves a lot of direct matching and lower order skills.

Overall the CBSE examination papers are heavily biased to rote-learning, do not test higher-order skills, and actively discourage students who try and display them.

Pakistan

As in India, there exist many examination boards in Pakistan. Each province, and often each district within a province, will have its own examination board. In addition to these regional boards, there are two national boards – the Aga Khan University Examination Board (AKU EB) and the Federal Board of Intermediate and Secondary Education – both of which also offer Secondary Certificate examinations.

These boards of intermediate and secondary education offer matriculation examinations at Grades 10 (the Secondary School Certificate [SSC]) and 12 (the Higher Secondary School Certificate [HSSC]). This study obtained papers from Azad Jammu and Kashmir, Sukkur, IBA Private School Sukkur, Hyderabad, and Larkana as examples of the regional boards.

The overall percentages for both sets of examinations are shown in the tables below.

SSC Grade X Examination (% by Skill Level)			
	English	Mathematics	Science
Level 1 Know	88	22	96
Level 2 Apply	12	78	4
Level 3 Reason	0	0	0

HSSC Grade XII Examination (% by Skill Level)			
	English	Mathematics	Science
Level 1 Know	91	13	93
Level 2 Apply	9	85	7
Level 3 Reason	0	2	0

Hidden within this there are large variations between the various boards. Even within Sindh, application in mathematics varies between 60 and 85 percent depending on the board and paper. Between regions there are also large variations in the balance of skills. This is shown in the table below.

SSC	Sukkur	Public School	Hyderabad	Larkana	Azad Jammu + Kashmir
English					
Know	87	na	96	89	80
Apply	13	na	4	11	20
Reason	0	na	0	0	0
Maths					
Know	23	40	14	17	17
Apply	77	60	85	82	83
Reason	0	0	1	1	0
Science					
Know	98	92	99	98	92
Apply	2	8	4	2	8
Reason	0	0	0	0	0
HSSC					
English					
Know	90	na	99	97	79
Apply	10	na	1	3	21
Reason	0	na	0	0	0
Maths					
Know	7	5	3	7	44
Apply	93	86	97	93	56
Reason	0	9	0	0	0
Science					
Know	95	93	91	92	95
Apply	5	7	9	8	5
Reason	0	0	0	0	0

Within question papers there are often different options that should be equivalent, but differ both in demand of the content and the task. There seems to be very little effort made in ensuring equivalence between boards, between subjects, over time, or even within papers.

The Federal Board of Intermediate and Secondary Education recently carried out a major review of the quality of outcomes for the HSSC from various boards. By correlating student performance on the HSSCs to student performance on university admissions tests,[17] the review showed significant variation between HSSC results and university admission test results. One of the purposes of the HSSC examination is to identify which candidates are ready for admission to university. Students are often still required to take a university admission test in a particular area of study and the two assessments should be highly correlated. The review does assume that the admission tests are selecting the right candidates, but this is a safe assumption because the admission tests were originally introduced and designed to select the right candidates (as universities believed that the

[17] Quality and Standardisation: A Twin Dilemma of Public Examinations at Higher Secondary School Level in Pakistan: FIBSE Review 2017

HSSC was not doing a good job). In the Sindh boards (those studied above), the correlations ranged from at best 0.54 through 0.46, 0.22, 0.19 down to 0.07. Khyber Pakhtunkhwa (KPK) had the lowest correlations which were between 0.22 to 0.01. The boards in the Punjab were better with a range of 0.66 to 0.50. The pan-Pakistan Boards (the Federal Board, Cambridge O Level and AKU-EB) ranged from 0.68 to 0.57.

This suggests that the majority of the HSSC papers studied here are failing to accurately predict which students should be going on to study at university.

Given the quality of the papers observed, this is not surprising. Overall the quality was very low, the question papers were poorly produced in terms of type-setting, and had many mistakes and unanswerable questions in them. In some of the non-English papers, the mistakes in language occasionally made it hard to understand the question and even some of the English papers contained grammatical mistakes. The analysis focuses on the 2016 papers, but to draw conclusions on things such as whether similar contexts and questions are being repeated and becoming predictable, papers going back several years were studied.

The learners are clearly intended to memorise huge tracts of knowledge and reproduce it on the page. Questions are not so much questions, but are prompts for candidates to regurgitate learnt answers. This is well illustrated by the opening question of an English paper.

> Q1 Answer **any five** of the following questions.
>
> i) State the dangers of marine fishing
>
> ii) What does everyone in China do?
>
> iii) For what purpose was the village decorated?
>
> iv) The Punjab is called the seat of learning? Why?
>
> v) The author climbed the steeple. Why?
>
> vi) What pleasure does the railway journey give to the poet?
>
> vii) Where is Nigeria situated?

All seven options are looking for reproduction of the information provided verbatim in the set texts. These are all impossible to answer correctly with no context other than given above or to score the available marks without knowing the answers given in the mark scheme. In many cases, this is taken to the extreme that the candidate must know the exact word used in the original text in order to distinguish synonymous answers. For example:

Tent pegging is another popular ….?

a) Show
b) Profession
c) Sport
d) Pass time

Or

The village people generally awake at ...

a) Midnight
b) Dawn
c) Sunrise
d) 4 .am

Bear in mind that both these examples come from the objective test and so no further context or quotes are given – the candidate is simply expected to know the exact right answer.

The English and science examinations are testing almost exclusively knowledge; obscure and mostly useless knowledge, at that. The content of the English papers is mostly concerned with specific details of the text (e.g., the exact wording used or incidental and often irrelevant facts that do not relate specifically to the message or meaning of the text - dates, ages of characters, number of lines in a poem, etc.).

Similarly, the science papers are testing almost pure recall of facts including some that are more properly the history of science than part of a useful scientific curriculum (e.g., *'When was the national science council set up?' 'What did Robert Brown discover in 1831?'*).

In these examinations, it would be possible to score high marks from pure rote memorisation and yet be completely functionally illiterate in English or science in any meaningful sense.

Mathematics is slightly better in that it tests mostly application of mathematical principles, but this is often in a highly structured format and it is open to debate how much requires genuine application and how much is rote learnt response.

As with the other countries, the Pakistan papers contain nominal composition exercises in English, but are structured to encourage regurgitation of learnt responses rather than creativity. The format of choice seems to be a written letter, either applying for a job, writing to a relative or friend, describing an event, or conveying some bad news. Moreover, the content of the letter can be readily transferred to other questions after replacing a few key details. As in the CBSE examples, the mark schemes are aimed mostly at procedural aspects and inclusion of stated facts.

The prompts asking students to write an essay are repetitive and broad. For example, variations of prompts asking students to write an essay titled, *'All that glitters …'* or *'Knowledge is power'*, occur several times across exam boards and years, as do other common themes such as, *'state of the education/examination system'* and *'the role of the media.'* It is not surprising that there exists a thriving industry of on-line providers of ready-made essays dealing with the most likely variations.

The outcome of all this is that these papers score very low on higher-order skills and most score low even on Level 2.

Alberta

In Alberta, assessment is the individual schools' responsibility and most of the assessment is carried out by the teacher following guidelines set by the Provincial Ministry of Education, Alberta Education.[18] There are also centrally provided Provincial Achievement Tests at Grades 6 and 9 to help support teacher judgments, encourage school improvement, and to allow school management, authorities, and the province to monitor performance and standards of learners. These tests are administered in May and June each year. Whilst the latest versions of these tests are confidential and so not available for study, Alberta Education released older versions of the tests which were used for our study.

The overall percentages for both sets of examinations are shown in the tables below.

Alberta Grade 6 Achievement Test (% by Skill Level)			
	English	Mathematics	Science
Level 1 Know	0	0	8
Level 2 Apply	88	90	62
Level 3 Reason	12	10	30

Alberta Grade 9 Achievement Test (% by Skill Level)			
	English	Mathematics	Science
Level 1 Know	0	0	0
Level 2 Apply	76	87	76
Level 3 Reason	24	13	24

The almost total absence of direct recall reflects the philosophy of the curriculum in Alberta. This is encapsulated by the Science Programme of Study statement:

> *"Children learn to inquire and solve problems in a variety of contexts. Each subject area within the elementary program provides a rich source of topics for developing questions, problems and issues that provide starting points for inquiry and problem solving. By engaging in the search for answers, solutions and decisions, students have a purpose for learning and an opportunity to develop concepts and skills within a meaningful context."*

[18] https://education.alberta.ca

The focus is very much on providing contexts within which concepts can be assessed, rather than assessing facts and recall. As the curriculum has a very limited number of assessable facts stipulated, the assessment needs to focus on application of concepts.

The tests are produced to high standards and showed no obvious errors or mistakes. None of the items studied were ambiguous or wrong and the distractors (the wrong answers in the multiple-choice items) are all plausibly wrong and aimed at expected misconceptions or mistakes.

As well as many items testing application of knowledge, there are items in each test that are clearly aimed at testing higher-order skills. The following example is taken from Grade 6 science.

> Paul wanted to investigate the movement of the Sun over the course of a day. He used a sundial to measure the length of a shadow once every hour from sunrise to sunset.
>
> Which of the following variables must be kept the same in order to obtain reliable data from this activity?
>
> A. Type of sundial and location of sundial
>
> B. Length of shadow and location of sundial
>
> C. Type of sundial and the times at which measurements are taken
>
> D. Length of shadow and the times at which measurements are taken

Similarly, in the same paper, learners are given a diagram of a crime scene and asked to make some logical deductions from it (e.g., '*Where did the culprit likely enter the house?*'). This requires the learner to analyse and evaluate information in the picture to answer correctly.

Mathematics also has items requiring learners to analyse and critique information, for example at Grade 6:

> Hannah wants to know if Grade 6 students in her school prefer skiing to snowboarding.
>
> Which of the following groups of students should Hannah survey?
>
> A. Students on the Grade 6 ski team
>
> B. All Grade 6 students in her school
>
> C. Students on the Grade 6 snowboard team
>
> D. Grade 6 students from the school's ski and snowboard club

Grade 9 mathematics also has similar items.

> Nina and Sarah observe that 6 of their 10 female classmates are shorter than 160 cm. Nina concludes that of the 410 students in their school, 246 are shorter than 160 cm. Sarah believes Nina's conclusion cannot be supported by her observation.
>
> Which of the following statements best supports Sarah's belief?
>
> A. Nina's survey sample contains only female students.
>
> B. Nina's probability calculation is incorrect.
>
> C. Nina did not use a proper questionnaire.
>
> D. Nina completed her survey too quickly.

The English Achievement Test also does a good job of testing whether learners understand text given as a stem to the questions. There were no examples found that could be answered by simple lifting of text or matching words without understanding the meaning. For example, an extract from 'The Secret of Nimh' is given and, as well as identifying the meaning of words and phrases in the specific context of the text, learners are expected to analyse and explain the purpose behind specific sections of the text. For example:

> The purpose of lines 1 to 8 is mainly to
>
> A. develop setting
>
> B. develop conflict
>
> C. characterize Mr. Ages
>
> D. characterize Mrs. Frisby

The texts given are intended to be unfamiliar to the learners and are taken from a wide range of age-appropriate material likely to be relevant to the learners. For example, the learners are given a Peanuts cartoon (a popular cartoon found in many newspapers featuring Charlie Brown and his child friends) where Lucy, one of the main characters, is caught day dreaming in a mathematics class. The learners are presented with the complete cartoon strip and asked a series of questions on it, including ones that require the learners to attribute motives to the character's actions.

In frame 6, the most likely reason that the girl repeats all of the numbers is that she is

A. making the problem clear to her classmates

B. stalling for time before giving her answer

C. showing off her knowledge of numbers

D. trying to solve the problem in her mind

Overall, the assessments from Alberta are well constructed and clearly show a link between the assessment and the aims of the curriculum and are designed to promote learning and assess it in a meaningful way.

Summary

There are a few good examples where the examinations attempt to assess higher-order skills, the Ugandan Primary Leaving Examinations and Nigerian NCEE especially, but overall higher-order skills are not assessed in the vast majority of assessment material from the developing world countries studied. In the worst cases, the assessments actively discourage higher-order skills, especially in the Indian and Pakistan papers. There is a very heavy focus on direct recall of knowledge, and in many cases, this knowledge is not useful or relevant.

The primary examinations in Nigeria and Uganda in mathematics and science seem to compare well with Alberta (although there is the caveat that the demand of the content being assessed has not been compared) and the expected benchmark, but the science examinations focus very much on direct recall of scientific facts. In India and Pakistan, high stakes assessment materials at this level were not available.

Figure 1 Comparison of primary (Grade 6) English examinations

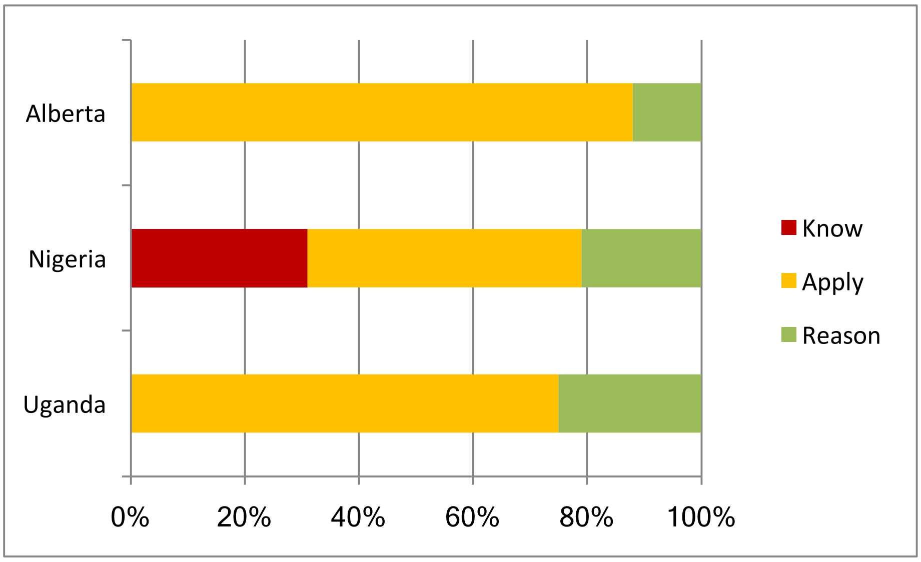


Figure 2 Comparison of primary (Grade 6) mathematics examinations

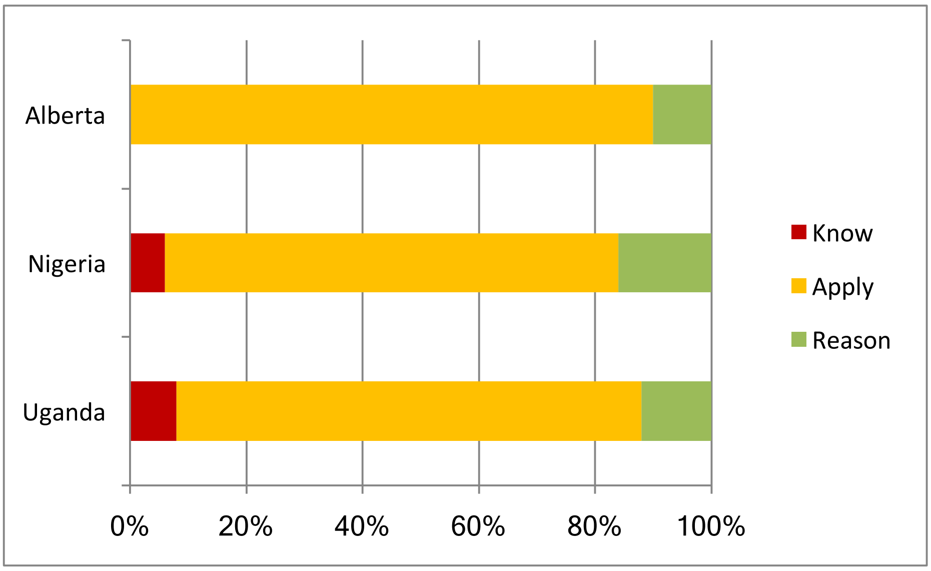
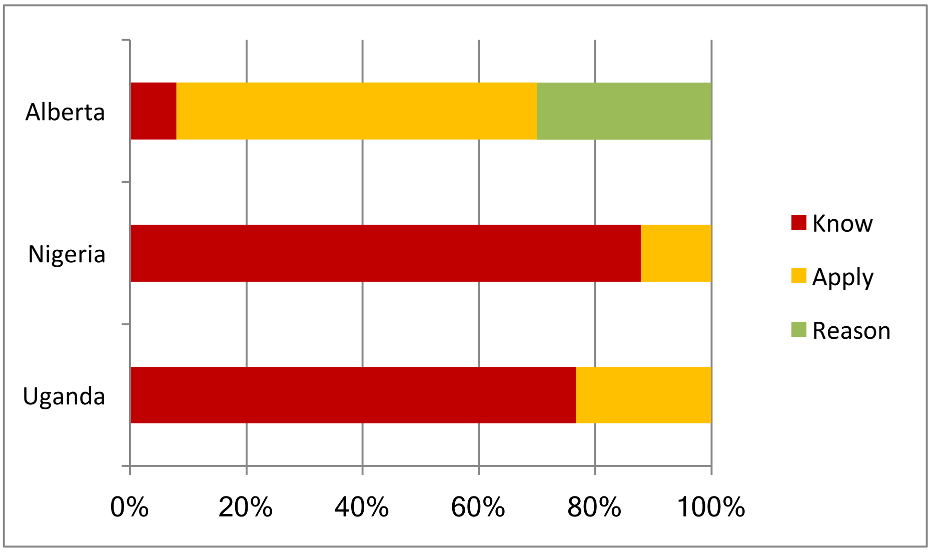


Figure 3 Comparison of primary (Grade 6) Science examinations



At Grades 9 and 10, the end of basic education in many countries, the proportion of higher-order skills assessed is much less than in Alberta, and the focus on direct recall much higher. The only exception is Uganda, although even there the science examination, as with the Grade 6 examination, has a strong focus on recall of scientific facts.

Figure 4 Comparison of lower secondary (Grade 9/10) English examinations

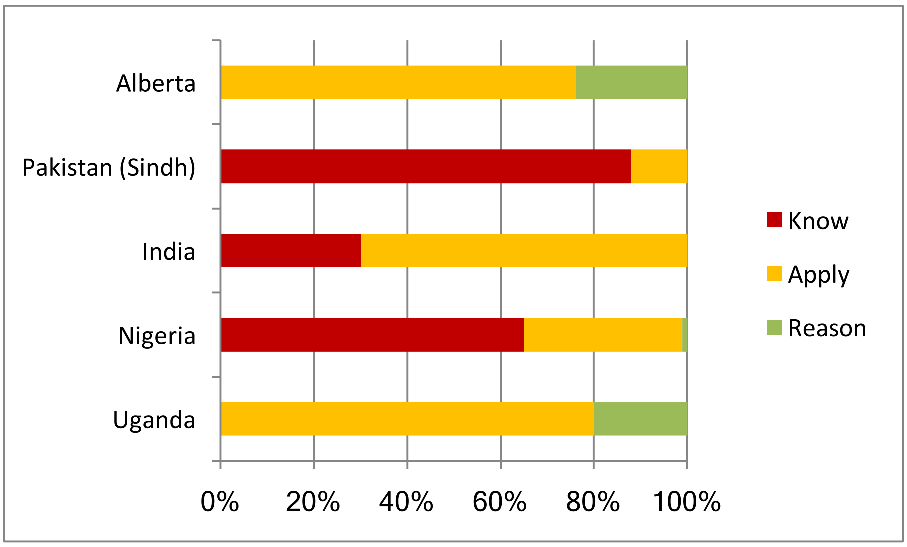


Figure 5 Comparison of lower secondary (Grade 9/10) mathematics examinations

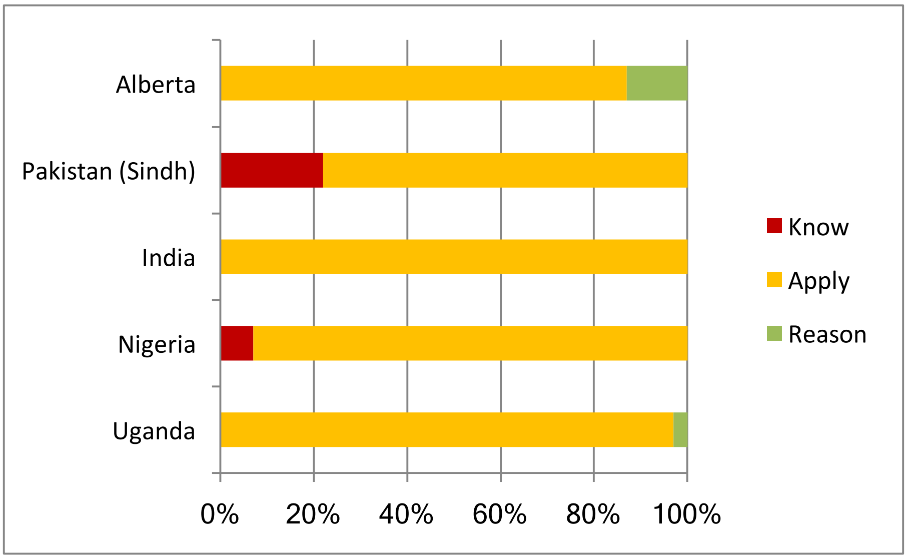
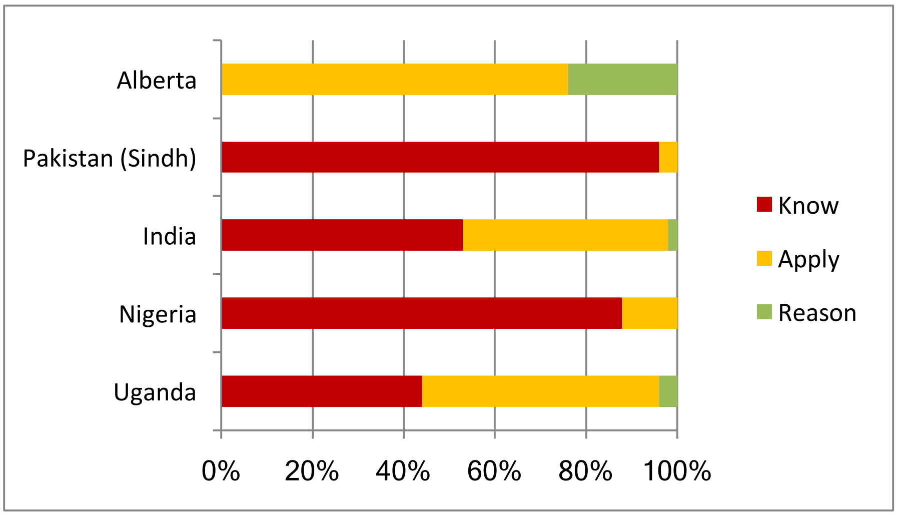


Figure 6 Comparison of lower secondary (Grade 9/10) science examinations



The focus on assessing high levels of recall, low levels of application, and higher-order skills in science is troubling as it implies that the science curricula are crammed with facts that need to be recalled, but that ways of scientific thinking and inquiry are being neglected. It might be that the curriculum does include these skills, but if the examinations do not test them, then it is unlikely that they will be taught. As the world becomes increasingly technological it is important that all students gain a basic understanding of the science and technology that surrounds them and can understand the debates that affect them in an informed way. Knowing a lot of dissociated scientific facts is not the same as being scientifically literate. By way of contrast, the Alberta Provincial Achievement Test's own science item classification lists 'recall' and 'understanding' together as a single classification of 'knowledge' and the majority of items test understanding, with the clear implication that knowledge without understanding is useless. The Alberta curriculum is designed for a rapidly changing world, the programme of study is designed to:

> *"…prepare students for life in a rapidly changing world—a world of expanding knowledge and technology in which new challenges and opportunities continually arise. Tomorrow's citizens will live in a changing environment in which increasingly complex questions and issues will need to be addressed. The decisions and actions of future citizens need to be based on an awareness and understanding of their world and on the ability to ask relevant questions, seek answers, define problems and find solutions."*

The assessment is also designed to encourage this aim and focus on assessing skills and concepts, rather than facts.

In mathematics, there is more focus on application and this hides an important issue. There is very little assessment of basic numeracy in the developing world assessments studied. Other than in the Ugandan PLE and NCEE there is almost no assessment of basic mathematical competencies and very little of the mathematics assessed is relevant outside of the classroom.

Similarly, the difficulty of the papers does not seem to be ramped in any sense (i.e., the lower demand questions at the start and the higher demand questions at the end). There may be an issue in that the setters are probably experienced mathematicians, possibly with a university background, and may not be capable of judging the demand required at these lower levels and seem to be producing items that are inappropriate for the candidates. The assessment, and the teaching that precedes it, needs to be at the appropriate level for the intended cohort.

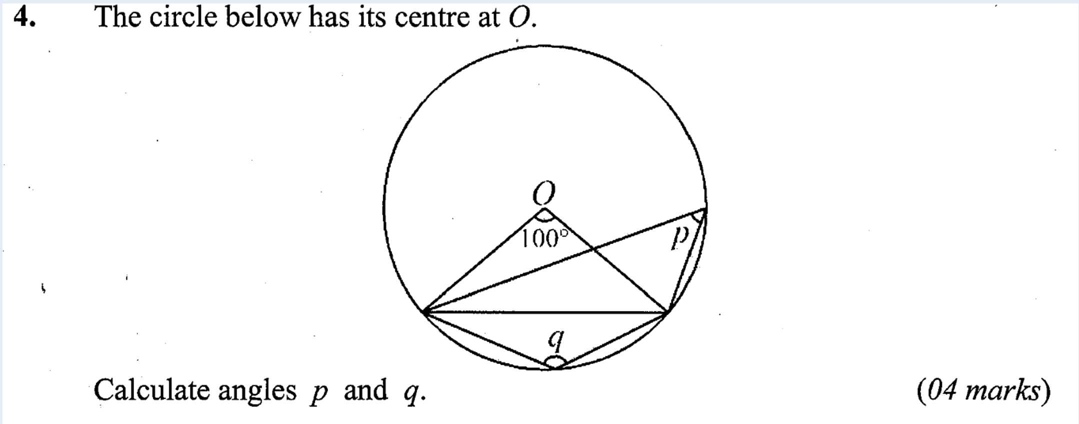
This focus on application of theoretical mathematics and failure to assess basic numeracy and mathematical skills is an issue especially when these skills are likely to be low. In the CBSE papers, based on the passing percentages, roughly 50 percent of learners are expected to get a C or below and are unlikely to be going on to higher education. It is difficult to identify which questions are aimed at these learners. In the West African Examinations Council, a pass would seem to be around 40 percent of the marks, but again it is difficult to define any content aimed at this lower boundary. The overall pass rate for WAEC in 2015 was 38 percent. This suggests that the majority of students are not being catered to by the content of these examinations and, by implication, the education system.

This study shows that it is possible to benchmark these systems against the IEA benchmark in terms of balance of skills, but it also highlights that it is a very different thing from benchmarking performance on these examinations. For several of these examinations, what is being taught and assessed is so divorced from what is expected to be taught and assessed in the international benchmarks, that making any meaningful comparison is impossible. It would be possible to score highly on one of the Pakistan papers, but fail to achieve the lowest benchmarks on IEA. Conversely, a student could score highly in PIRLS or TIMSS and fail to pass one of the CBSE or Pakistan Board Papers. The huge amount of very precise, extraneous recall required in these systems means that a significant part of the curriculum cannot be mapped onto any other benchmark system. The systems demonstrated by the papers from CBSE and Pakistan do not match in any meaningful way what is considered a necessary or good education in any of the available benchmarks.

Even with the higher quality assessments, for example the Ugandan papers or the NCEE, the presence of so much weighting on recall of required knowledge would require close curriculum mapping before trying to match outcomes with any benchmark.

Implications for the curriculum and classroom practice

The assessment materials suggest there is a large disconnect between what is being assessed and the skills and knowledge needed by learners leaving school. Some of the examinations do assess skills that have wider relevance, for example the Ugandan PLE and Nigerian NCEE assess mathematical competencies that are generally useful (e.g., simple addition, multiplication, money, percentages, interest rates, etc.), but most of the other examinations focus very much on traditional mathematics such as shown below.



4. The circle below has its centre at O.

Calculate angles p and q. (*04 marks*)

This is not bad in itself, but it must be remembered that very few of the learners studying for these examinations, which are school leaving examinations, will be going on to further study. Compare this to a more contextual question:

The exchange rates in a bank are as follows:

1 US dollars ($) = Ug. sh 3,400

1 British Pound Sterling (£) = Ug. sh 4,600

1 Kenya shilling (K.sh) = Ug. sh 35

(a) Convert Ug. Sh 1,840,000 to British Pound Sterling *(02 marks)*

(b) If a set of chairs costs $700, find the equivalent cost of the chairs in Kenya shillings *(03 marks)*

This is a big issue as it implies that much of what is being assessed (and therefore taught in schools) is not directly relevant to most learners when they leave school.

The assessment materials suggest that a lot of useless knowledge is being assessed, and by implication taught, especially in the Pakistan examples. This concern that the assessment is driving what is taught and over-riding good teaching practice, and any intended curriculum, is not a theoretical concern, but a well-documented one.[19] Many of the pieces of knowledge expected in the examinations are trivial, for example learners are expected to

[19] Adnan, U., Mahmood , M.A., Impact of Public Examination on Teaching of English: A Washback Perspective Journal of Education and Practice Vol.5, No.2, 2014

know (without any additional context) the exact price a Persian rug cost in one of the books they must study as part of the reading list.

How much was the cost of the Persian carpet?

a) 1432 Rs
b) 1332 Rs
c) 1532 Rs
d) 1632 Rs.

They are also meant to know the speed Mrs Oakentubb was travelling in her car in an obscure play studied in the Sindh secondary system (20 Minutes with Mrs Oakentubb), as well as numerous dates, ages, and other incidental details. This knowledge is not literature; this is just useless detail. They are not expected to know anything substantial about the texts or analyse or evaluate them in any way. These examinations, and by implication what is being taught in the classroom, have lost touch with why learners should study literature and the reason why set reading lists are on the curriculum. The purpose of studying English literature texts is not for learners to acquire useless trivia from minor plays that have long been forgotten outside the BISE curriculum. But these examinations seem to have become detached from the skills and learning that they should be assessing, and instead have driven the learning process in damaging ways, hijacking the curriculum and focussing on surface trivia.

The level and vocabulary being tested is also inappropriate. In countries struggling with basic literacy, why are obscure and archaic words being tested? What use is it to a child at any of these educational stages in any context, let alone the developing world, to know the meaning of *‘pugilistic encounters’*, *‘sanguinary’*, or *‘hoary’*? In the context where these definitions were asked, it was not even possible to work them out from the surrounding context, so they were pure recall. The purpose seems more to demonstrate the setter’s erudition rather than to assess any meaningful measure of literacy.

Similarly, many have retained texts and pieces of knowledge that are extremely dated. Whilst tent pegging, a cavalry sport where a rider galloping at speed tries to remove a tent peg from the ground with a lance, is still practiced around the world, it is unlikely to be a sport familiar to many candidates and surely better and more relevant examples could be found? Set texts and contexts should include some modern material and the more dated material should only be retained if it has real, educational merit. This is not just confined to the core subjects; one Geography paper expects students to know how many fisheries were active in Baluchistan in 1947.

In science, there is a strong focus on very specific scientific knowledge, but scientific literacy and the ability to judge scientific statements or make informed decisions is lacking. The CBSE tries to introduce these with questions designed to assess attitudes in science, but these questions are poorly designed and do little to support this initiative.

Much of the science knowledge is also of dubious use, especially when so much can be looked up. Why is it important to know which scientist studied the Alkaptonuria genetic disorder in detail? Or which part of the cell was discovered in 1831 by Robert Brown?

In some of the examinations it was not clear whether what was being assessed was on the curriculum. For example, many of the Nigerian English items seemed to be assessing history or general knowledge rather than vocabulary. In a Pakistan HSC Botany paper, there is a question about jackals and predation that is completely unrelated to plants.

In one statistics paper, what is the question, '*Give the brief history of statistics*' testing? Similarly, in the same examination, what is the question below testing?

> The word Statistics is used in
>
> a) Many sense
> b) Four sense
> c) Two sense
> d) All of these

What do the setters of these papers think candidates should be learning? Looking at most of these examination papers it is unclear whether the setters are following any sort of test specification or setting guidance linked to clear learning outcomes, or whether these learning outcomes are of any use in the real world.

Many of the curricula seem to be in most cases designed for students going on to further study, and not the majority of learners, and even then, it is not clear how much of the content taught is still relevant and likely to be used after leaving school. Many of the important basic skills, including language literacy, scientific literacy, mathematical literacy, and numeracy seem to be mostly absent. Uganda and the Nigerian Federal system do seem to value these skills, but the Nigerian states', Indian, and Pakistan systems seem to be clinging on to assessing content that seems to have remained unchanged for decades.

Language demand

The language demands of the examination papers in India, Pakistan, and Nigeria seem high, even when compared to the Alberta papers and other examination papers from English speaking developed nations. This is even though the majority of the candidates will have English as a second- or third-language. Sometimes the language used is obscure and would be better phrased more directly. For example:

> A student is studying the properties of acetic acid in his school laboratory. List two physical and two chemical properties which he must observe and note in his record book.

Would be better phrased as:

> State **two** physical and **two** chemical properties of acetic acid.

In many of the comprehension passages more could be done to edit the texts to make them accessible to the target audiences without reducing the level of comprehension demand. For example, in one of the Junior Secondary Nigerian examinations (aimed at 13 year olds) the opening paragraph discusses the parotid gland, sub-lingual gland, and sub-mandibular gland, as well as having many polysyllabic, technical medical terms that are not needed to answer any of the following questions. These could all have been edited out to simply say '*Saliva is produced by glands in the mouth'*. This is much more appropriate to the intended age and language level.

In contrast, the Ugandan papers tend to use direct and simple language much more appropriate to the target audience (e.g., '*Calculate the amount of money Hajati has after 3 years'* or '*How is the sugar weighed according to the story?').*

Bad test items

One of the biggest issues with all these papers is the proportion of bad test items. There are numerous examples where it is unclear what is being asked or what the expected answer is. It is common in the multiple-choice items to find more than one correct answer or occasionally no correct answer.

> The chemical mostly used in the preparation of most of the soaps we use is
>
> a) Sodium chloride
> ***b) Potassium hydroxide***
> c) Sodium hydroxide
> d) Potassium chloride

The first three are all commonly used in soap manufacture and it is hard to see how a candidate can be expected to judge the 'most' used in 'most' soaps. Similarly, with:

> We eat food for ….
>
> a) To reduce hunger
> b) To get energy
> c) To enjoy test
> d) To live

only one answer is clearly wrong (unless *test* is a mis-spelling of *taste*, which is not impossible given the mistakes in spelling and grammar scattered throughout the paper this was taken from; in which case, it is even harder to judge the right answer.)

Or:

> ………………… is represented of balanced diet.
>
> a) Fish
> b) Apple
> c) Egg
> d) Milk

is meaningless as it stands, as is the example below.

> A student is testing water to know which is best for cleansing purposes with soaps. He would find that the cleansing action of soaps is best when he uses water obtained from
>
> (a) rain
>
> (b) tap
>
> ***(c) hand pump***
>
> (d) pond

Why is water from a hand pump best for cleansing with soap? What is being tested here? Is there a single right answer? Scientifically, rain water is probably the chemically softest and so might lather best, but there is no way to judge the quality of the others from the information given.

Some just leave you wondering:

> If the clouds on the sky how many chances of rain
>
> a) 0.005
> b) 0.05
> c) 0.5
> d) None of these

The examples above are from Pakistan and the presence of so many bad items probably goes a long way to explaining the poor correlations between the Pakistan HSSC results and the university entrance tests, but similar items can be found in all the other countries. The examples below are from the Nigeria SSCE papers.

> Then suddenly, his (the young doctor's) eyes caught a woman on a nearby
>
> a. stretcher
> b. **bed**
> c. wheelchair
> d. bench

With no additional context to differentiate the answers, it is impossible to say which is most appropriate; all are possible and make sense within the text (except possibly wheelchair which would be ‘*...**in** a wheelchair*’).

> The boy’s parents … (a)raised (b) uttered (c) breathed ***(d) heaved …*** a sigh of relief.

Both ‘*breathed’* and ‘*heaved’* a sigh of relief are correct and used in every-day English. *‘Heaved a sigh of relief’* is a more idiomatic expression, but not necessarily more correct. The example below, from the Indian CBSE, is another one where it is hard to justify the scoring answer as being the only or even most correct answer.

> The patient… (a) was dying ***(b) had died*** (c) could die (d) died …. before the doctor came.

The Ugandan examinations and the Nigerian NCEE have fewer outright bad items and seem to have better quality control, but still have occasional examples of poorly constructed questions that are unclear or break good assessment practices. For example:

> Which of the two following reactions is SN1 type?

With only two options to choose from, a candidate who does not know has a good (50:50) chance of guessing the right answer.

In all the systems studied there seems to have been less thought put into how the questions are marked than their development. How the questions are to be marked and what are acceptable answers is as important as the development of the question itself. Often it seems that an arbitrary number of marks is assigned and either all awarded or not. For example:

> Josephine obtained 95% in a test marked out of 80 marks. How many marks did she score out of 80? [4 marks]

There is only one real answer (76) and it is hard to see how this merits four marks given that the response is either right or wrong, and there is not real scope for method marks or partially correct answers.

The number of outright errors, as well as instances of poor assessment practices found in many of these papers, suggests that if these papers are broadly representative of the developing world, many learners are not being fairly assessed. These errors are likely to be serious issues with using national high stakes examinations to monitor educational outcomes due to these tests unreliability and poor validity.

The examinations are also probably failing in the primary purpose of selecting the right learners for further education or employment. As evidenced in the Pakistan correlation study or the Seychelles example noted in the RISE Working Paper 16/010, high stakes examinations often appear to be failing to identify the correct learners, wasting resources and generating unfairness.

Positive aspects

It is obvious from the Ugandan materials, and from various statements made by Ugandan politicians, that this is a country trying to improve its curriculum and assessment system.

> *'… Education and Sports Minister, Ms Janet Kataha Museveni said the public has continued to judge the education success only in terms of examination results and grades attained by learners instead of what has been taught and learnt. … She said: "I am glad that most of you are educated experts and I implore you to critically look at our education system and identify the gaps and propose solutions to our education system especially assessment and examination which influence class room practice," she said.'*[20]

The papers studied show there is variable progress being made in Uganda on this front with the primary papers leading the way. At the higher levels, the landscape is more varied with some subjects showing higher-order skills than other subjects. It is possible that the higher-level papers will catch up as the reform makes its way through the system. The General Paper obviously exists to try to assess higher-order skills. Similarly, the presence of subjects such as entrepreneurship in the curriculum, with a significant proportion of the assessment being on higher-order skills, demonstrates that the system is trying to adapt and evolve to changing needs.

In Nigeria, the NCEE papers also show some good quality items and attempt to include higher level skills, although the other papers lag far behind.

In India, in addition to the state papers reviewed above, papers were also supplied from a private supplier, Educational Initiatives PVT Ltd (EI Pvt Ltd), which were designed to give diagnostic feedback on student learning. These showed high quality items including items that test higher-order skills. These items require students to demonstrate reading and comprehension skills and cannot be answered easily without understanding the text (unlike many of the CBSE questions which just require matching and no understanding). For example, one gives an edited poster advertising the rights of consumers under the Consumer Protection Act 1986 and asks questions (such as, *'Who is the narrator?'*) that can only be answered by reading and understanding the whole text. Compare this to the question on who is the narrator from the CBSE papers that require students to simply know the narrator after recognising the text. The same company, EI Pvt Ltd., that produced those papers also conducted research that replicated many of the findings here and argues for

[20] http://www.monitor.co.ug/News/National/Uganda--examination--system-education-Janet-Museveni-/688334-3958996-t8usm7z/index.html

gradual introduction of more items not taken directly from text books to gradually move students and teachers away from direct learning.[21]

In Pakistan, the Aga Khan University Examination Board (AKU-EB)[22] was set up at the request of schools to address the problems identified in this research - rote learning, poorly written examinations - as well as general corruption and malpractice. They continue to conduct research into how to improve both the content and delivery of examinations and to share that knowledge with examination boards in Pakistan. The level of quality of the AKU-EB papers and balance of skills is generally much higher than the public board papers.

For example, compare this text analysis question from AKU-EB Grade 12 English paper (below) with the comprehension questions found in the public board papers.

Explain the author's purpose in writing this article. Give at least ONE example from your surroundings to support or negate the author's view. (3 marks)

Stimulus/ Stem:

Adapted from 'Fat Shaming is Not Community Service' by FarazTalat

Possible Answer

To persuade: The author's purpose is to persuade the readers that fat shaming is not community service. It is a kind of bullying.

OR

To inform: The author wants to inform that we should not assume that obese people are lazy / undisciplined/ indolent and that we should not judge a person if he/ she is unable to lose weight.

Example: phrasing will vary as per candidate's thoughts

Checking Hints:

2 marks for describing the author's purpose (Give 1 mark if the student mentions 'to inform / to persuade but does not mention the complete description / evidences)

1 mark for mentioning ONE valid example which supports or negates the author

The AKU-EB comprehension paper requires learners to read the text given in the examination rather than pre-learning rote-responses or doing trivial matching exercises. This

[21] Reforming Board Exams for Learning with Understanding. Raghav Rohatgi & Pranav Kothari Educational Initiatives Pvt. Ltd. (India) https://secure.hbcse.tifr.res.in/epi6/papers/Strand-3-posters/epi6_P-46_Raghav%20Rohatgi%20&%20Pranav%20Kothari.pdf

[22] The author should note that he is a non-stipendiary, executive board member for the AKU-EB

genuinely requires students to engage with the text and analyse it in a meaningful way and attribute a motive to the author. This is higher-order thinking. It also allows for more open-ended responses and for the students to come up with more than one correct answer. Students are required to select examples from the text to justify their choice and it is clearly not looking for a rote learnt answer. Interestingly, the AKU-EB categorises this item as understanding rather than application, but a good case could be made for this being a higher-order skill.

In contrast to the public board papers, rather than trying to assess useless fragments of information, the whole paper is trying to test whether students can understand a text and can analyse it in a deeper way. The paper has questions designed to assess whether students can distinguish between what is clearly stated and what is implied, whether they can understand the tone and how the author supports his/her opinions.

> “Individuals born with metabolic silver-spoons in their mouths, snicker at people struggling with their weight problem, much like rich kids who roll their eyes at poor people struggling to pay their bills.”
>
> Explain the meaning of the given phrase. Why does the author compare obesity with poverty? (3 marks)

This again requires genuine explanation and students need to understand the phrases, how the author has structured the sentence, and to compare and link the sections dealing with weight and poverty and draw valid conclusions from it. The AKU-EB does not re-use comprehension items and so there is no incentive for students to rote-learn predicted answers; they can only answer these questions by showing comprehension skills.

Recommendations

Improvements to assessment materials

Whilst there is evidence of good practice in the studied papers, the examination boards would benefit from more capacity building in assessment practices, especially in writing items that test higher-order skills accurately. In some countries, this will be building on an existing base, but in others that base would need to be built almost from scratch.

All the countries studied need to review their assessment practices, evaluate what they currently do against best practice, and set up a cycle of continuous improvement. Beyond the obvious examples of poor questions shown above, there were numerous examples from all the developing world examinations of poor assessment practices (e.g., where the question and mark scheme did not align, either in terms of demand or expected response, or where multiple routes to a final mark meant different learners effectively sitting different examinations). The various bodies should be encouraged to review and overhaul their quality control procedures to ensure that the assessment materials produced are fit for purpose and contain no errors. As many of these examinations act as gateways to further education or employment opportunities, it is important that they select the learners as accurately as possible.

In India and Pakistan, and to a lesser extent in Uganda and Nigeria, there needs to be a move to develop assessments that discourage rote-learning and instead focus on ensuring students can demonstrate the skills they will need when they move beyond school, whether that is for further education or employment. The items in the examination need to become a lot less predictable and include material that is not taken directly from text books.

As more developing countries seek to introduce assessments to monitor learning, for example the Indian Certificate of Secondary Education council (ICSE) is introducing additional examinations at Grades 5 and 8 to enable monitoring of learners' progress[23], it is important that those assessments accurately and validly measure learning and encourage good teaching, rather than further entrenching poor practices.

Impact on curriculum and pedagogy

To alter the learning outcomes effectively, any changes in assessment need to be mirrored in changes in curriculum and in pedagogy (and vice versa). In many cases, the curriculum (as assessed by these tests) seems to be defined by the text books used and this practice either needs to be abandoned or the text books improved and up-dated at the same time as the assessment materials are revised.

If there is a greater proportion (or even some) higher-order skills in the examination, then teachers will need to spend more time teaching these skills. This will mean that a lot of content will need to be removed from the current curriculum, but given the debatable utility of much of that current content, this should not be a problem. It should be noted that changing the exams does not guarantee that teachers will know how to teach these skills. The incentives provided by a high stakes examination may be properly aligned with desired outcomes, but that does not guarantee the capability to deliver them.

In Pakistan, it is not clear if the examination papers studied were following the latest Federal national curriculum, but given the content, it is unlikely. A good starting point would be for these boards to look to reviewing their curricula, possibly against the national curriculum, and abandon some of the older text books.

In the systems that are already trying to inculcate higher-order skills in their learners, the assessments need to drive effective teaching of these skills and not encourage 'short cuts' or undermine teaching of these skills.

There is a significant issue in that much of the teaching force might need extensive training to teach these skills as it requires a very different pedagogic approach. It is doubly hard if the teachers themselves do not have sufficient capacity in these higher-order skills having never been taught them whilst they were studying.

Indeed, these skills may not be taught during initial teacher training, and this would need to be addressed as they require a different teaching approach. It may well be that these skills

[23] http://www.hindustantimes.com/education/icse-to-hold-board-exam-for-class-5-and-8-students-from-2018-arathoon/story-DXgHi3mSw7nQHypAhTPcfK.html

are being taught, but then suppressed once newly qualified teachers are in the schools as the demand for more-examination based, rote learning overrides any training they have had. It would be interesting to investigate the content of teacher training courses and how this links to the skills studied here.

This paper has not had time to investigate the readiness for the teaching profession to adapt to new assessment regimes, but it is likely they will struggle. In Uganda, there is a clear desire in the various statements from politicians and ministry officials to change teaching practice, but the fact that learner outcomes remain low suggests that this is a hard issue to address. In Nigeria, there is a clear difference in standard between the NCEE examinations and the other examinations. It is likely that this is largely influenced by the quality of the federal schools and teachers and it is unclear how well teachers in state schools would cope with a change in curriculum and assessment. There is also the issue that having learnt higher-order skills for the NCEE, the system then reverts to focusing more on lower level skills.

In India and Pakistan, it is unclear how well teachers who are used to only didactic teaching of rote-learnt material would transition to a different teaching approach.

Proposal for a public good

This paper attempts to establish a framework for comparing examination papers across countries, or subnational regions within a country, along a single dimension; the extent to which exams are testing for higher-order skills. This is only one aspect of measuring educational outcomes, but an important one. Hopefully, creating the conditions for the objective analysis of the quality of examinations, even on only a single dimension, allows policy makers to have more meaningful discussions about policy objectives and improves the research base for reform agendas. There needs to be better data about the quality of the examinations as they are such important drivers of behaviours in education systems and this data needs to be available to those involved in educational reform to allow evidence-based decisions to be made.

Those wishing to improve educational outcomes need an effective accountability framework.[24] It is important that the examinations embedded in the system are of sufficient quality to be a coherent part of that system, rather than introducing further incoherence into the system. Examinations lie at a crucial nexus in the education system and provide both information and motivation in an accountability system. The ability to understand and analyse examinations is a key part to understanding the wider system in order to change it effectively.

Of course, a full accounting of an examination's quality can, and should, include at least a few other dimensions. Assuming this can be done with a reasonable degree of rigor, it is likely that policy makers and examination boards would find a wider-toolkit that compares examination instruments cross-nationally quite useful; no large-scale attempt at this is known to the author. SABER, the Systems Approach to Better Education Results[25], has diagnostic toolkits for looking at assessment, but these focus on processes and the policy frameworks

[24] http://www.riseprogramme.org/files/RISE_WP-005_Pritchett.pdf

[25] http://saber.worldbank.org/index.cfm

within which assessments sit, rather than the quality of the measurement instruments themselves. Among reform-minded policy makers, such comparative analyses could help increase the political salience needed to initiate change, while technical members of an exam board could draw on the materials directly. Most of the existing literature on examinations focuses on the political economy of their high stakes consequences and issues of cheating and corruption, which is of little direct use for practitioners.

Some of the key issues that would need to be included in such an effort are detailed below:

- **Curriculum alignment** - An examination not properly aligned to the national curriculum will be working at cross purposes to stated policy intent. Students who spend the bulk of their primary school years traveling down one learning progression, may find themselves preparing in their final two years for an examination that tracks something different. The meaning of the word “alignment” and how to measure it in this context, however, is not immediately straightforward. Some of the dimensions to be considered include mapping content domains (and subdomains), establishing a method for quantifying alignment and overlap (by exam item or domain), the number of grade levels covered by the exam, and more.

- **Curriculum relevance** - As well as the examination being aligned with the curriculum, the curriculum needs to be well aligned with the needs of the learners in terms of content difficulty, breadth vs. depth, current learning levels, abstract vs. practical learning, etc. Very few of the examples studied here focus on assessing functional or practical skills or content that has relevance outside the class room or for further study, even though most of the cohort sitting the examinations will not being going on to further study and should be learning skills that are relevant and useful to them.

- **Validity and reliability** - The examinations must test what they are claiming to test and do it reliably. The correlation data from Pakistan suggests that many of the papers set by the public boards are not measuring the skills that they should be, or at least not doing so accurately and reliably. The fluctuations on pass rates seen in many of these examinations suggest that they are not providing good year-on-year stability needed to make accurate comparisons over time. Similarly, this study suggests that the issue of where in the cohort distribution examinations provide good discrimination needs to be studied as the early indication is that they focus on discriminating at the higher-ability end (based on the questions seen) and possibly provide poor discrimination where the bulk of the intended cohort actually lies.

One useful check on the quality of an examination would be the public availability of anonymised student outcomes data, which would enable researchers and policy makers to confirm whether exams are working as intended. Examination boards genuinely interested in improving the quality and performance of their tests should put these data into the public domain and make it available for research. The case of the Higher Secondary School Certificate in Pakistan is instructive: if matriculation exams are insufficiently correlated with future academic success, their fairness to students must be called into question.

Appendix A: Why bad assessment is bad for education – case study from Pakistan

Extract from RISE Working Paper 16/010 - The good, the bad, and the ugly - testing as a key part of the education ecosystem. The full text can be accessed at:

http://www.riseprogramme.org/content/rise-working-paper-16010-good-bad-and-ugly-testing-key-part-education-ecosystem

Nobody would argue that Pakistan's education system faces huge challenges[26] 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. Part of the problem is the examination system in Pakistan and this is true right the way from the primary examination 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.[27,28,29] 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 disconnect between what should have been taught and assessed 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 repeats followed by the question and the answer, then gets the class to repeat this back and forth chant 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.*

[26] 11th Education for All Global Monitoring Report http://en.unesco.org/gem-report/

[27] Rehmani, A. Impact of Public Examination System on Teaching and Learning in Pakistan International Biannual Newsletter ANTRIEP, 8 (2) Pp.3-7 2003

[28] Mirza, M., Nosheen M., Masood N. 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. 1999

[29] Kamrani, S. Future of Pakistan in respect of Education, Islamabad: Aziz Publishers 2011

> *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 repeat the scoring phrases exactly 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, and yet despite 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.

And if nobody has gone through a good education system and has only experienced the same deficiencies, it becomes very difficult for them 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 currency than learning itself.

As Adnan and Mahmood[30] summarise it:

> *"Need for success in HSSC 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 to a larger or lesser degree are seen as high stakes. 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.

Whilst this replacement of learning by marks and certification is not universal it is highly entrenched as can be seen by remarks made by various members of parliament when exposed in the fake degree scandal.

"A degree is a degree! Whether fake or genuine, it's a degree! It makes no difference!" - Baluchistan province chief minister Nawab Aslam Raisani.[31] The learning does not matter, just the façade; education reduced to a cargo cult.

[30] Adnan, U., Mahmood, M.A., Impact of Public Examination on Teaching of English: A Washback Perspective Journal of Education and Practice Vol.5, No.2, 2014

[31] Reported in The Dawn June 30 2010

References

Anderson, L. and Krathwohl, D. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Pearson.

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

Bethell, G. (2016). *Mathematics Education in Sub-Saharan Africa: Status, Challenges, and Opportunities.* World Bank, Washington, DC. Available at: https://openknowledge.worldbank.org/handle/10986/25289 License: CC BY 3.0 IGO.

Biggs, J. and Collis, K. (1982). *Evaluating the Quality of Learning: The SOLO Taxonomy*. New York: Academic Press.

Braun, H., Kanjee, A., Bettinger, E. and Kremer, M. (2006). *Improving Education Through Assessment, Innovation, and Evaluation*. [online] American Academy of Arts and Sciences. Available at: https://www.amacad.org/publications/braun.pdf

Burdett, N. (2017). *RISE Working Paper 16/010 - The good, the bad, and the ugly - testing as a key part of the education ecosystem | www.riseprogramme.org*. [online] Riseprogramme.org. Available at: http://www.riseprogramme.org/content/rise-working-paper-16010-good-bad-and-ugly-testing-key-part-education-ecosystem.

Chaudhuri, S. (2017). *ICSE board makes Sanskrit, yoga and performing arts compulsory subjects*. [online] http://www.hindustantimes.com/. Available at: http://www.hindustantimes.com/education/icse-to-hold-board-exam-for-class-5-and-8-students-from-2018-arathoon/story-DXgHi3mSw7nQHypAhTPcfK.html

Daily Monitor. (2017). *Uganda proposes overhauling examination system*. [online] Available at: http://www.monitor.co.ug/News/National/Uganda--examination--system-education-Janet-Museveni-/688334-3958996-t8usm7z/index.html

Griffin, P. and Care, E. (2015). *Assessment and Teaching of 21st Century Skills: Methods and Approach*. Springer Netherlands.

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

Kellaghan, T. and Greaney, V. (2004). *Assessing Student Learning in Africa*. Washington, DC: World Bank. Available at: https://openknowledge.worldbank.org/handle/10986/14910 License: CC BY 3.0 IGO.

Malik, I., Sarwar, M. and Imran, A. (2017). *Quality and Standardization: A Twin-Dilemma of Public Examinations at Higher Secondary School Level in Pakistan*. FBISE REVIEW. [online] Islamabad: Federal Board of Intermediate and Secondary Education. Available at: https://www.fbise.edu.pk/Downloads/QUALITY%20AND%20STANDARDIZATION.pdf

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.

OECD. (2017). *The case for 21st-century learning*. [online] Available at: http://www.oecd.org/general/thecasefor21st-centurylearning.html

Raghav Rohatgi & Pranav Kothari Educational Initiatives Pvt. Ltd. (India). *Reforming Board Exams for Learning with Understanding*. Available at: https://secure.hbcse.tifr.res.in/epi6/papers/Strand-3-posters/epi6_P-46_Raghav%20Rohatgi%20&%20Pranav%20Kothari.pdf

Rehmani, A. (2003). Impact of Public Examination System on Teaching and Learning in Pakistan. *International Biannual Newsletter ANTRIEP*, 8 (2) pp.3-7.

Suto, I. and Eccles, H. (2014). The Cambridge approach to 21st Century Skills: definitions, development and dilemmas for assessment. In: *IAEA Conference*.

Uganda National Examinations Board (2017). *STATEMENT ON RELEASE OF 2016 UCE EXAMINATION RESULTS*. [online] Available at: http://uneb.ac.ug/downloads/2016_UCE_RELEASE_STATEMENT.pdf

UNESCO (2017). *Global Education Monitoring Report*. Accountability in education: Meeting Our Commitments. [online] Paris: UNESCO. Available at: http://unesdoc.unesco.org/images/0025/002593/259338e.pdf