HEART
HEALTH & EDUCATION ADVICE & RESOURCE TEAM

UKaid
from the British people

UGANDA'S ASSESSMENT SYSTEM: A ROAD-MAP FOR ENHANCING ASSESSMENT IN EDUCATION

Reg Allen, Phil Elks, Rachel Outhred and Pierre Varly[1]

14 September 2016

[1] Reg Allen is the author of the main report and Annexes A, C and E. Pierre Varly and Reg Allen authored Annex B. Rachel Outhred and Phil Elks wrote Annex D.

Acknowledgements

The authors would like to thank all those many people who so generously gave us their time and contributions, and the schools that welcomed us.

The findings, interpretations, and conclusions expressed are ours and do not in any way represent the views of the Government of Uganda or its agencies.

Any errors, omissions, inaccuracies or misinterpretations are our responsibility.

We appreciate the help the Uganda National Examinations Board (UNEB) gave in providing us with data, under agreed protocols, and with reports. As required under these protocols, we have sought comments from UNEB on a pre-publication draft of our report of the results of our analyses and have responded to any comments or corrections from UNEB about the methods, results and conclusions of the report and the extent to which our report's use of the provided data is consistent with the mutually agreed purpose.

We also appreciate the help the National Curriculum Development Centre (NCDC) and the Directorate of Education Standards (DES) gave us in providing a wide range of useful reports, together with the extensive input provided by their staff members in a range of meetings.

Finally, we would like to extend a particular thank you to Florence Buyinza from UNEB, Gertrude Namubiru from NCDC, Annet Mugisha from the Teacher, Instructor Education and Training Department (TIET) and Joseph Lubwama from DES, who worked with our team during the school based fieldwork reported in Annex D of this report.

This assessment is being carried out by HEART (Health & Education Advice & Resource Team).

The team leader is Reg Allen and the project manager is Rachel Outhred. For further information contact rachel.outhred@opml.co.uk .
The contact point for the client is Ed Barnett E-Barnett@dfid.gov.uk. The client reference number for the project is [7825-A0826-PO 40096612].

HEART
Level 3, Clarendon House
52 Cornmarket Street
Oxford, OX1 3HJ
United Kingdom

Tel +44 (0) 1865 207 333
Email info@heart-resources.org
consultancy@heart-resources.org
Web www.heart-resources.org

Executive summary

Assessment in education in Uganda takes many forms: for example, the Primary Leaving Examination (PLE) and Uganda Certificate of Education (UCE) examinations, the National Assessment of Progress in Education (NAPE) and other system-wide tests, practice examinations used in schools, formal and informal classroom assessment, and so on. Uganda's assessment system is a complex and interconnected web of all these forms of assessment, together with the actions of institutions and people involved directly and indirectly with the processes and results of assessment.

This report provides a road-map describing how Uganda can develop a robust, resilient, self-improving and adaptive assessment system, grounded in Uganda's culture, history and context.

Such an assessment system will be able to produce the results Uganda needs and wants both now and in the future under changing economic, social, political and external circumstances.

The current assessment system does not meet all of these requirements.

Our report describes the current system under seven headlined descriptions:

1. In practice, little priority is given in classrooms to matters that are not closely aligned to the demands of examinations.
2. Standardised tests of literacy and numeracy do not show significant and continuing upward trends and the link between using these tests and changes in policy and practices is not clear.
3. Teacher absences from classrooms are a fundamental constraint on classroom learning and assessment processes.
4. There are significant regional (e.g. rural/urban) and social differences in participation and attainment.
5. Much of the knowledge and skills learners need today and for their futures has little or no role in formal tests and examinations.
6. Ensuring integrity and authenticity in high-stakes assessments presents significant challenges.
7. Many parents look for ways to ensure their children get the high scores needed to get into 'good' schools for the next stage of their education.

We find that the current assessment system's key results, which reflect the features described under these headlines and their interaction, include:

- *too many students who are successful in school do not learn enough of the skills and knowledge aligned with the current and future needs of employment and the further development of Uganda's economy;*
- *public discourse about education focuses more on scores than on what is learned and its alignment with the needs of individual and community success now and in the future; and*
- *reports by agencies often do not lead to effective action – at central, district and local level.*

While no-one set out to design the totality of the current assessment system, it functions as if it had been carefully designed to achieve exactly the results it achieves now. The design and implementation of change requires understanding current realities, and understanding why and how they interact to produce the results they do.

We found that Uganda's assessment system in 2016 has a focus on examinations designed for selection and we identified consequent negative impacts on teaching and learning. The Education

Policy Review Commission, chaired by Professor Senteza Kajubi, made similar observations in 1989. The underlying stability of the current assessment system continues despite various reports identifying problems and proposing technical solutions, and despite well-written policy and curriculum documents.

An enhanced assessment system will be one with:

- *clear, explicit and published standards for the processes and products of the agencies with responsibilities within the assessment system;*
- *open, honest and transparent reporting against these standards;*
- *a climate in which reporting against standards is used for continuous improvement rather than blame or sanctions; and*
- *positive and negative incentives institutionalising and reinforcing appropriate practices of individuals and agencies.*

The steps involved in making changes under the enhanced system we propose are:

1. agreement amongst stakeholders that a particular feature is really important and is worth improving or changing (a 'problem');
2. using data from reports about agencies' processes and products to develop a shared understanding of the many reasons and causes of this feature and its interaction with other aspects of the assessment system;
3. using this shared understanding to agree on some feasible and affordable interventions that will have a significant effect;
4. implementing these interventions;
5. reviewing the impact of these interventions; and
6. adaptation, adjustment and refinement (or identification of a deeper 'problem').

Most of the problems that exist under Uganda's current assessment system are what are sometimes called 'wicked problems': so-called because they are complex, multi-faceted, and difficult to solve, and because an effort to solve one aspect of the problem may lead to other, unexpected, difficulties.

As an example of such challenges, changing current teacher practices in assessment will not be achieved by simply issuing instructions or by organising a widespread programme of teacher professional development. What teachers currently do in assessment in Ugandan classrooms reflects not only their individual abilities and capacities but also, predominantly, the complex set of multiple interacting factors and influences we have sought to describe in this report. Effective action to change teacher assessment practices will need to be built on a clear understanding of the current situation, including the factors leading to teacher absence, agreement amongst stakeholders about the importance and value of changes, and the development of actions that can be resourced and will be cost-effective.

Changing examinations and other forms of assessment to include assessment of unrehearsed problems[2], higher level concepts, creativity and collaboration will not happen solely with a change in the techniques and technicalities of assessment. It can, however, happen over time, through the agreement and active contribution and engagement of all stakeholders (including parents, teachers and the community), through well-designed interventions or changes, through a review of their effectiveness and through further adaptation.

[2] For example, questions that require students to apply knowledge to an unfamiliar situation or questions requiring students to develop a process that they have not previously practised.

Designing interventions or changes in assessment practices in schools must be based on a clear understanding of why schools use the practice tests they currently use and how these meet needs of schools, teachers and parents.

Given such an understanding and appropriate resourcing, a set of interventions might consist of a planned programme that coordinates incremental changes over time in the content and format of examinations, including:

- the professional development of teacher leaders at the district level;
- the provision of appropriate practice tests across grade levels at a price competitive with the price of commercial tests;
- helping community stakeholders, such as parents and employers, to be actively and effectively involved in the development and successful implementation of changes;
- gathering systematic information about what happens in practice, whether intended or unintended, as changes are implemented; and
- using this information as the basis for developing and implementing adaptations of changes to make them more effective.

There are no single actions or techniques that will bring about change overnight. By themselves, individual changes to elements of assessment/examinations, importing ideas and processes from elsewhere, or replacing one kind of assessment with another will not lead to an adaptive, resilient self-improving system, one working within its culture, context and resources. There are, however, significant strengths within the current system that can be built on in the development of the standards-based approach we propose. These strengths include the demonstrated capacity of the National Curriculum Development Centre (NCDC) and the Uganda National Examinations Board (UNEB) to carry out careful reviews of their work, and the work of the Directorate of Education Standards (DES) in developing, applying and reporting on school performance against standards.

Consistent with the basic principle of continuous improvement based on review and monitoring, the elements in our proposed road-map do not constitute a set of fixed steps, but rather a set of actions that are influenced by and build on the capacities, understanding and agreement among stakeholders developed by earlier steps. This is a process that starts now and continues over the next five to ten years, a process in which technical developments are ***consequences***, not a starting point. Technical developments are made when of an agency has identified opportunities for improvement of its processes and products against a set of standards that have been developed collaboratively and agreed by government. In turn, later reports show the extent to which such technical developments have in fact led to the expected improvement in the quality and effectiveness of processes and products.

While the scope of our review was restricted to assessment at the primary and lower secondary levels, these form part of an overall assessment system, in which features at one level affect and are affected by features at another level – higher or lower. Implementation of the processes set out in this road-map must therefore take these interactions into account.

Table of contents

List of figures and tables

List of abbreviations

BTVET	Business, technical and vocational education
DES	Directorate of Education Standards
DFID	Department for International Development
EGRA	Early Grade Reading Assessment
ESSP	Education Sector Strategic Plan
GPE	Global Partnership for Education
HEART	Health and Education Advice and Resource Team
ICT	Information and communication technology
IRT	Item response theory
KIIs	Key informant interviews
MENA	Middle East and North Africa
MLA	Monitoring of learning achievement
MoES	Ministry of Education and Sports
NAPE	National Assessment of Progress in Education
NCDC	National Curriculum Development Centre
OPM	Oxford Policy Management
PISA	Program for International Student Assessment
PLE	Primary Leaving Examination
SACMEQ	Southern African Consortium for Monitoring Education Quality
TIET	Teacher, Instructor Education and Training Department
UACE	Uganda Advanced Certificate of Education
UBTEB	Uganda Business &Technical Examinations Board
UCE	Uganda Certificate of Education
UNEB	Uganda National Examinations Board
USAID	US Agency for International Development

1 What this report does

The purpose of this report is to set out the main features of the proposed road-map for enhancing Uganda's assessment system by:

- identifying the senses in which the term 'assessment' is used in this report
- defining the main elements of Uganda's assessment system at the primary and lower secondary levels;
- showing that key 'headlined' descriptions of Uganda's assessment system, when looked at in isolation, separated from their context and culture, can be found in other developing and developed economies;
- explaining how and why the proposed road-map is based on the idea of an overall, systems approach, one that is integrated, strategic and positioned in Uganda's history, context and culture (rather than a set of unrelated actions, each focused on one of these headline features in isolation), through:
  - *describing how Uganda's current assessment system is better understood as a complex and interconnected web*
  - *describing the key results of the current assessment system*;
- setting out the main elements of the proposed road-map by describing:
  - *the destination*
  - *current strengths that can be built on*
  - *elements along the way*;
- describing how the road-map responds to the current assessment system's headline features and key results; and
- discussing implementation, governance and risks

Five annexes to this report describe what we learned in this study from:

A. analysing UCE and PLE sample data sets:
   i. student-level data
   ii. 300 marked scripts in each of four PLE examinations;
B. NAPE reports and data sets;
C. analysis of PLE and UCE examination papers and commercial practice tests;
D. classroom observations; and
E. reports of public discussions of the purposes, products and outcomes of Uganda's assessment system.

These annexes include short descriptions of methods and, where appropriate, reference to the alignment of the curriculum with assessments. Technical tables and graphs used in the preparation of Annexes A and B appear in a separate document (an additional appendix called 'Roadmap tables and graphs') attached to this report.

In developing this road-map we have drawn heavily on a systems design perspective [3], one in which we have analysed Uganda's assessment system as a complex and interconnected web of

[3] See Allen (2006) for a systems perspective on assessment and certification; see Stroh (2015) for a general account of systems thinking for social change; see the working papers of Michael Woolcock and others at Harvard's Center for

agencies and individuals connected by ideas, values and priorities communicated through actions. We have sought to begin to build some understanding of what *actually* happens (as opposed to intentions or aspirations) as the basis for developing proposals about ways in which Uganda can develop effective and sustainable enhancements in its assessment system.

A focus on the main features of the current assessment system is essential to developing effective enhancements but brings with it the risk that our clear and simple descriptions will be understood as fault-finding or taken as occasions for blaming institutions or individuals within agencies. We have tried to reduce this risk in the way we have set out the material in this report, while also avoiding the risk of obscuring the main characteristics of the situation with too many caveats and complications.

While the scope of our review was restricted to assessment at the primary and lower secondary levels, these form part of an overall assessment system, in which features at one level affect and are affected by features at another level, higher or lower. Some references to assessment at these levels are therefore necessary.

Our report does not make detailed recommendations about a set of technical developments or capacity-building, or specific changes in forms of assessment that should be implemented. Instead we develop proposals for the development of a continuous improvement process in which such changes will be designed, resourced, implemented, reviewed and adapted as part of building an assessment system that will meet Uganda's needs now and in the future.

International Development https://www.hks.harvard.edu/centers/cid/ for papers discussing problem-driven iterative adaptation, such as Andrews, Pritchett and Woolcock (2016).

2 The functions and purposes of assessment

Assessment in education has many forms and many purposes and functions.

The function that matters most is the way assessment drives and shapes what is taught and learned.

It is really important that the assessment system makes sure that learning in schools includes all the learning it should, and that *most attention* is paid to the *most important* learning, not the learning that is easiest to assess.

Both the official and in-practice purposes of forms of assessment (examinations, tests, quizzes, checklists, observation schedules, performance assessments and so on) can vary widely and sometimes a given form of assessment can have multiple and conflicting purposes.

The purposes of assessment include:

- reaching decisions about a student's demonstrated achievement in a course of study (for certification[4]);
- reaching decisions about a student's ability or aptitude (for admission, classification or placement);
- reaching decisions about how best to help a student to learn ('formative' assessment for planning instruction); and
- producing sufficiently reliable measurements of a latent trait (for accountability – i.e. assessing the 'performance' of teachers, schools and systems).

The main function of assessment, however, is the way in which in practice it drives, for better or for worse, both instruction (what teachers do) and learning (what students learn). That is, whether or not an assessment used in schools is for certification, for competitive selection into further education or a career, whether or not an assessment is 'of, as or for learning', what that assessment does is to drive instruction and learning. Its impact on learning may be strong or weak, may deepen or trivialise, may widen or narrow – but an impact it will have.

The central purpose of assessment in schools is therefore, we believe, to ensure that students, in practice, learn all that they should learn and learn most what it is most important that they learn. In other words, assessment is there to make sure that learning in schools includes *all the learning* that it ought to, and to make sure that *most attention* is paid to *the most important learning*, not the learning that is easiest to assess.

The other purposes of assessment in schools are means to an end, not ends in themselves. For example, accountability of schools through various forms of performance measures (often based on standardised test results) are based on the idea that this is a good way to improve, not restrict, learning. Certification gives students qualifications that provide other people with an official record of what they know and can do. The high-stakes nature of certification means it strengthens the impact of assessment on learning. Assessment for selection purposes raises the stakes for students and strengthens its impact on learning.

The higher the stakes in an assessment the greater its impact on learning, and hence the more important it is that we ensure that this impact is positive, directed towards encouraging more learning by more learners of the body of knowledge and skills that is most important for their futures.

[4] See the notes on the terminology used in this report in Addendum A.

3 Assessment in Uganda: A range of forms of assessment, together with multiple agencies and individuals involved in assessment processes and products, constitute the Ugandan assessment system

There are many forms of assessment in education in Uganda, and many agencies and people are involved.

Planning change requires a focus on realities, not on finding fault or on intentions. There is no simple path from policy statement to change.

It is important to look at Uganda's assessment system as a whole, one that operates across all levels of education.

3.1 Assessment: What happens

In primary and lower secondary education in Uganda, assessment takes a range of forms – examinations, practice examinations, commercial tests, locally devised tests, classroom observations, standardised tests and household-based surveys.

Considered together, these forms of assessment involve a wide range of institutions and people.

There are the autonomous/semi-autonomous statutory agencies:

- UNEB, which has a statutory remit[5] to set, administer and mark examinations and standardised tests, supervise and collect results of continuous assessment, issue certificates and publish examination papers, results and reports (including reports of its research); and
- NCDC, which has a statutory remit[6] to develop and evaluate syllabuses and their implementation, and to work with examining bodies on devising, testing and evaluating examination questions and methods of examining students.

There are also agencies that are part of the Ministry of Education and Sports (MoES), including:

- DES, which sets standards for the quality of education in Uganda, and reports against these standards at school, district and national level. DES implements school-based monitoring of learning achievement (MLA) that has a diagnostic focus, supporting schools to make the required improvements (Department of Education Standards 2014); and
- the Teacher, Instructor Education and Training Department (TIET), which supports, regulates and promotes quality teacher, tutor and instructor education.

There are also non-government agencies, such as Uwezo, which assesses and reports on levels of basic literacy and numeracy.

The Early Grade Reading Assessment (EGRA) has been developed through programmes supported by the US Agency for International Development (USAID) and now, with funding from the Global Partnership for Education (GPE), it is being conducted by UNEB.

There are schools, public and private, which, in terms of assessment, select and provide tests and practice examinations for learning and reporting purposes, and which gain status and reputation from published tables of their students' results in examinations.

[5] See http://uneb.ac.ug.
[6] See http://www.ncdc.go.ug.

There are students, who directly experience assessment in its many forms and who learn from it and through it.

There are teachers, who have a direct role in what happens in classrooms, including classroom assessment in its various forms, who shape their teaching to suit the content and skills required by examinations, and who are affected by the public reputation of their school.

Finally, there are members of the community, who form and exchange views about the success of schools and education generally through public discussions of examination results.

3.2 Uganda's assessment system: The importance of a focus on realities

Uganda's assessment system is each and every one of the forms of assessment listed above together with the actions of the many institutions and people involved directly and indirectly with the processes and results (outputs and outcomes) of assessment. This system forms a complex and interconnected web, one discussed in more detail later.

Actions and practices, rather than intentions, are what really matter. While curriculum and syllabuses state aspirations and expectations, there is no simple and automatic path from these documents to the teaching and learning that happens in classrooms. What happens in assessment – especially the assessments that 'count', the ones where the results are seen as important by students, teachers, schools, parents and the community – shapes and directs teaching and learning. At the same time, such assessment also reflects taken-for-granted (culturally and historically mediated) assumptions by teachers, schools and parents about what students can and should learn. This is a complex relationship, one not correctly characterised by the idea of 'teaching to the test'.

3.3 Uganda's assessment system operates across all levels of education

The scope of our review was restricted to assessment at the primary and lower secondary levels. However, these form part of an overall assessment system, in which features at one level affect and are affected by features at another level – higher or lower.

For example, assessment at the lower secondary level is influenced by assessment at the upper secondary level, where results in the most prestigious subjects select students for entry to competitive university places. In turn, this has an impact on participation patterns in business, technical and vocational education (BTVET) – patterns that are low and that mainly involve lower achieving students.

In BTVET content and assessment, we have seen examples of courses at diploma level with a clear focus on workplace skills and competencies that also include elements that appear to be there essentially for the purpose of allowing graduates to seek entry to university programmes. The strategic plan for BTVET (Ministry of Education and Sports 2011) calls for a paradigm shift so that the BTVET system will 'emerge from an educational sub-sector into a comprehensive system of skills development for employment, enhanced productivity and growth.' This plan characterises Uganda's current assessment system as focused on educational certificates rather than on the skills and competencies that are relevant in the labour market.

Another example of such assessment system interactions having significant impact across multiple levels of education may be found in reports we heard of formal interviews for selection for entry to some kindergartens in Kampala. This sort of interaction can lead to the situation found in in Japan, where there are establishments that coach pre-schoolers and their mothers on how to gain entry to a prestigious kindergarten, so that students will be on track to enter the 'right' schools so as to gain entry to the 'right' universities.

The many ways in which features of the assessment system interlock and interact with values, attitudes, aspirations and priorities of parents, students, teachers and schools mean that change is not simply a matter of setting and promulgating official policies and making technical changes in examinations, tests and performance management systems.

To ensure a clear focus on the main, key and central features of Uganda's current assessment system, we have not included reference to assessment for students with special needs (or high needs). We recognise the challenges involved in doing this well (as well as its importance) and are aware of the provisions made in this area by UNEB. Ensuring that the assessment system drives high quality learning for all students is important for many reasons, including the backwash effect on the quality of what happens in classrooms. We also recognise the importance of the early childhood sector, while noting that this is outside our current scope.

4 Key individual features of Uganda's current assessment system

'*If the standards and quality of education in Uganda are to improve, all stakeholders must work in open, transparent partnerships focused on the single shared aim of improving learners' experiences and achievements.'* (MoES Annual Report 2012, p.5)

The seven headlined descriptions of Uganda's current assessment system we set out below are features that, taken in isolation, can be seen in other countries.

These are complex and difficult problems, ones with no simple, single solutions that can be imported, ready-made, from countries with a different culture, history and context.

In the context of assessment, the aim of improving learners' achievement requires a focus on what, in practice, students actually learn and how assessment, as it actually happens, drives this learning. This is a focus on realities, not on intentions or aspirations.

Describing such realities simply and briefly is a task for 'stylised facts': generalisations based on observations and analysis. Such description is not criticism, fault-finding or blame. These statements are only 'headlines' – the complete story is, of course, much more nuanced and complicated. As is clear from the more detailed discussion in the annexes, these headlined descriptions are over-simplifications about what happens in Uganda – but they are useful for the purposes of this report.

The headlined descriptions are useful in two ways. First, they provide a sharp focus on essential features that a road-map for evolutionary enhancement of Uganda's assessment system must take into account. Secondly, in each case, such headline features, when taken in isolation, can be seen in a range of other countries, including both developing and developed economies. This does not mean that a remedy proposed, adopted or actually implemented in these economies for each of these individual features can be successfully imported by itself into Uganda, but it does mean that we should recognise that these are difficult problems, ones with no simple, single solutions. These complex problems cannot be treated separately from their situation. They must be seen together, not one by one. They must be treated in their situation in their context of place, time, history and culture. The systems approach described later takes this wider perspective, recognising that Uganda is a unique country, with its own culture and history.

We describe seven headlined descriptions in total, starting with two focused on the relationships of examinations and tests with what happens in classrooms and developments in policies and practices:

- In practice, little priority is given in classrooms to matters that are not closely aligned to the demands of examinations.
- Standardised tests of literacy and numeracy do not show significant and continuing upward trends, and the link between using these tests and changes in policy and practices is not clear.

These two headlined descriptions are followed by four that describe different perspectives on practices:

- Teacher absences from classrooms are a fundamental constraint on classroom learning and assessment processes.
- There are significant regional (e.g. rural/urban) and social differences in participation and attainment.

- Much of the knowledge and skills learners need today and for their futures has little or no role in formal tests and examinations.
- Ensuring integrity and authenticity in high-stakes assessments presents significant challenges.

The last headlined description focuses on the impact of parents on practices:

- Many parents look for ways to ensure their children get the high scores needed to get into 'good' schools for the next stage of their education

4.1 In practice, little priority is given in classrooms to matters that are not closely aligned to the demands of examinations

Annexes A, C and D provide an account of what we learned through analysing PLE and UCE examinations, and the relationships between these and the practice examinations used in classrooms, and what we learned from classroom observations – our own and others. These show that the generalisation in this headline, although it does not capture the whole story, has sufficient validity for the purposes of developing this road-map.

It was clear to us that there is in Uganda a common and widespread standard practice of schools giving all students (even those in lower age groups) a beginning, middle and end of term examination style test. We have found little indication of these assessments being used to inform teaching through critical reflections and through providing individualised feedback to students.

We were able to review some examples of commercially available tests used in Grades 4 and 7. (P4 and P7). These show unmistakable signs that they are designed to reproduce the sorts of items to be found in the PLE examinations.

In practice, school accountability (see Annex E) is publicly expressed through examination results. Success in PLE and UCE examinations is not only high-stakes for students, it is also high-stakes for schools and for teachers.

Taken in isolation, this headline could be taken from a report about:

- education in Kenya
    - *' …the backwash effect on the curriculum … the teachers cannot engage all the methods of instruction to attract the learners' interest in class, selection of the content is determined by examinability and integration of life skills in the various subjects was also ignored or rushed over'. (Chang'ach et al 2012)*
- education in East Asia
    - *'… the overwhelming pressure of public examinations that overshadows the entire school system; the overemphasis on examination scores throughout the school system; the didactic pedagogy which precludes innovations and creativity.' (Cheng 2014)*
    - *'… much of the schooling in Vietnam was (and still is) more rote "chalk and talk", where the teacher writes information on a chalk-board, and the students copy and memorise it. No questions, no discussion. A one-way flow of information.' (Napier and Hoang 2013); and*

- education in the USA
    - *'… tests … [have] had negative effects on teaching and learning, especially for poor and minority students ...[teachers] provide practice on exercises that substantially match the format and content of their state's end-of-year… tests. These exercises often depart substantially from best instructional practice … Principals and district administrators encourage this practice. They introduce interim assessments that largely mirror the end-of-year tests rather than model the kinds of performance intended by the standards...' (Resnick and Berger 2010).*

4.2 Standardised tests of literacy and numeracy do not show significant and continuing upward trends, and the link between using these tests and changes in policy and practices is not clear

Annex B provides a more detailed and nuanced account of what we learned through analysing available NAPE information.

Uganda has conducted NAPE assessments in primary Grades 3 and 6 in most years since 1996. There have also been NAPE assessments in some secondary years since 2008.

According to Acana (2006, p.2):

> '… the information from these tests is supposed to serve the following purposes, among others: generate accurate information on what pupils know and can do in various curricular areas, provide guidelines for improving instruction and learning evaluate the effectiveness of inputs and processes that are employed in the educational activities, provide guidelines on variables that affect achievement. However, the information is not yet being optimally used by education policy makers and practitioners'.

The 'theory of change' behind large-scale sample testing programmes, such as NAPE, appears to be that the existence of data about students' learning will identify the need for changes in policies and practices, may illuminate the sorts of changes that are required, and will show that these changes have been effective.

It is clear from reports such as the most recent NAPE tests (Uganda National Examinations Board 2015) and a World Bank study of data up to 2012 (Najjumba and Marshall 2013) that there is no significant and continuing upward trend in NAPE results. This suggests that the information has not been converted into widespread and effective changes in teacher and school practices.

The details discussed in Annex B illuminate factors associated with the tests that may contribute to this:

- test construction may lead to the inclusion of items that do not add sufficient additional information[7] and the omission of items that test higher order skills;
- the linking across years does not seem to provide sufficient precision to track trends over time accurately;
- sampling does not generally appear to be structured to obtain information that can track changes in district performance (or some other grouping where some policy changes have been implemented);

[7] In the technical language of psychometrics, *redundant* items. See http://www.rasch.org/rmt/rmt143a.htm for a brief summary.

- published reports of students' test performance in NAPE generally do not seem to provide clear indications of the implications for specific changes in teaching practices;
- it is not clear that the tests and their administration are designed in terms of their capacity to illuminate the effectiveness of, for example, changes in policies or practices at a nationwide or district level; and
- technical developments may not be implemented for a range of reasons.

There are likely to be more complex factors outside the tests and reports themselves that contribute to information from NAPE generally not being frequently used to design, develop and implement effective changes in policies and practices. The systems approach taken in this report represents an understanding of this complexity and a recognition of the need to work within it.

There are suggestions for greater strategic engagement between policy-makers and assessment managers (Elks 2016).

Technical improvements to NAPE tests – for example, the use of the techniques of modern psychometrics in item development, test construction and analyses – are better seen not as an end in themselves but as a *consequence* of developing clear and explicit statements of the expected impact of NAPE (and other standardised tests) on student learning in Uganda, together with ways of demonstrating this impact.

The lack of a clear impact of standardised testing on changing levels of student performance is familiar across the world. For example:

- in Australia
  - 'results in the PISA [Program for International Student Assessment] have, to put it kindly, not shown any upward trend in standards or in reducing equity differences, with actual declines in some areas'
  - 'the results of standardised tests have essentially flat-lined. If anything, there are suggestions that the testing programme and the use made of results leads to a focus on students close to the margin between success and failure, at the expense of the very high and the very low achievers. It has been argued from longitudinal studies that "literacy and numeracy scores have stagnated or fallen since the 1970s – despite a doubling of resources"; and
- in the Middle East and North Africa (MENA)
  - 'in the classroom, the results from early grade reading and maths assessments reveal that MENA children are lacking the required foundational reading and maths skills. Likewise, students in MENA countries fall behind those in most other countries in basic reading, maths and science, as measured by the Trends in International Mathematics and Science Study (TIMSS) and the 2011 and 2012 PISA test scores… The majority of fourth- and eighth-grade students in MENA countries participating in these tests scored 'below low' in maths and science.' *(Brixi, Lust and Woolcock, Trust, voice and incentives: learning from local success stories in service delivery in the Middle East and North Africa 2015, p. 46)*

4.3 Teacher absences from classrooms are a fundamental constraint on classroom learning and assessment processes

It is a basic assumption that effective classroom learning requires the presence in the classroom of a teacher and the use of class time for instruction and classroom assessment. Any discussion,

therefore, of options for making classroom assessment more effective and giving it greater weight must, before considering any other matters (such as professional development, data collection, integrity and validity), take into account the extent to which teachers are present in classrooms and class time is used for instruction.

The following table shows the extent of teachers' absence from classrooms, the limited instruction time (as a proportion of the school day) and the significant variation by region and school type in Uganda.

Table 1: Rates of teacher absences from classrooms in Uganda

	Uganda	Central	East	**Kampala**	North	West	Public	Private	Rural	Urban
School absence rate	**23.6%**	21.8%	26.0%	**10.8%**	34.0%	17.2%	26.7%	14.2%	27.8%	15.6%
Classroom absence rate	**52.7%**	46.6%	59.7%	**38.1%**	68.9%	41.7%	56.8%	40.4%	56.5%	45.6%
Classroom teaching time (ToT)	**3h19m**	3h44m	2h46m	**4h39m**	2h01m	4h07m	2h58m	4h18m	3h04m	3h59m
Minimum knowledge among teachers	**19.6%**	24.0%	23.9%	**18.9%**	14.7%	13.7%	19.9%	18.7%	17.2%	24.4%
Student-teacher ratio	**40.1**	28.5	50.8	**19.5**	58.1	32.7	46.4	19.8	42.7	32.2
Share of Students with textbook	**4.6%**	7.7%	3.8%	**1.4%**	0.8%	5.0%	5.7%	1.3%	4.6%	4.5%
Teaching Equipment availability	**94.5%**	95.0%	89.3%	**97.6%**	98.6%	96.1%	94.2%	95.5%	93.9%	96.5%
Infrastructure availability	**55.3%**	53.9%	47.5%	**65.8%**	66.7%	56.1%	60.0%	40.2%	53.7%	60.2%

Source: This dataset is provided courtesy of the Economic Policy Research Institute (EPRC) which did the deeper analysis of education service indicators using data from the World Bank's Service Delivery Indicators (SDI) survey of 2013.

This situation is not unfamiliar elsewhere:

- in Nigeria
  - 'it is currently estimated that about 20% of the teaching workforce in government primary schools are absent on a given work day … It was recently reported that a top official ... went around public primary schools in Uyo capital city; and found many teachers (and in some cases even head teachers and their deputies) absent from work without permission.' *(Ejere 2010, p. 115)*;
- in Sri Lanka and India
  - 'In Sri Lanka, for example, teachers without patronage must teach in rural, remote areas without the opportunity to transfer to urban areas….This phenomenon [ghost teaching sessions], which is common in India, occurs when teachers do not teach their classes but are still registered in time-keeping books and receive their full salary.' *(Transparency International 2011, pp 6-7)*; and

- in Kenya, the following table shows high rates of teachers' absence from classrooms, and the limited time spent teaching compared with the scheduled teaching day.

Table 2: Rates of teacher absences from classrooms in Kenya

Indicator	All	Public	Private	Difference (% point)		Urban Public	Rural Public	Difference (% point)
	(1)	(2)	(3)	(4)		(5)	(6)	(7)
Absence from school	15.5%	16.4%	13.7%	-2.7%		13.7%	17.2%	-3.5%
Absence from classroom	42.2%	47.3%	30.7%	-16.7% ***		42.6%	48.8%	-6.1%
Time spent Teaching	2 h 40 min	2 h 19 min	3 h 28 min	1 h 9 min		2 h 38 min	2 h 14 min	24 min
Proportion of lesson spent teaching	81.4%	77.9%	89.3%	11.4%		82.1%	76.6%	5.6%
Scheduled teaching day	5 h 42 min	5 h 35 min	5 h 54 min	18 mins		5 h 33 min	5 h 37 min	-4 min
Classrooms with pupils but no teacher	30.8%	36.0%	19.2%	-16.8% ***		35.9%	36.4%	-0.6%
Observations	306	239	67			66	173	

Notes: Weighted means using sampling weight. Results based on observations from 2,960 teachers in 306 schools. Data collapsed at the school level. Diff in column 4 (7) is differences in means between private and public (urban public and rural public). Superscript (*) denotes the difference, using standard errors clustered at the school level, is significant at the ***1%, **5% and *10% significance level.
Source: Martin and Pimhidzai, Service Delivery Indicators education, health - Kenya 2013, p. 34

Within these fundamental constraints, there are, of course, further factors affecting the nature and value of classroom assessment, factors that reflect the complex interactions within Uganda's assessment system described in section 5.

Annex D describes the results of the initial classroom observations conducted as part of this study. The results of this study suggest that many schools:

- implement extensive and intensive testing, to varying extents, and that often this involves using scarce resources to purchase commercial tests; and
- systematically plan and report (if sometimes in a very limited way) on student achievement through classroom assessment, but assessments often focus on recall or copying skills. The results of both informal and formal assessments do not seem to impact sufficiently on teaching and learning in classrooms.

The small number of schools visited and the limited time for the classroom observations phase of the project, however, limited the richness of the data collected and reported in Annex D.

We propose that data be collected in a larger sample of schools during the first year of implementation of the proposed road-map, to capture the full range of heterogeneity in approaches to assessment across schools in Uganda, and to refine and develop the road-map. This larger collection should have both a quantitative element across a larger number of schools to collect prevalence data while also using qualitative studies of a smaller number of schools to gain essential insights into, and a deeper understanding of, the factors that influence assessment practices.

Annex C describes the priorities for learning found in commercial practice tests – which, we understand (see also Annex D), are often used as a basis for the learning needed to pass examinations, for classroom assessment and for reporting to parents (Guloba, Wokadala and Bategeka 2010).

4.4 There are significant regional (e.g. rural/urban) and social differences in participation and attainment

'Overall, comparisons by location provide few surprises. The very large gaps in P3 and P6, while alarming from an equity and policy perspective, are not surprising when compared to research and assessment results in other countries (also see the NAPE reports for Uganda). There is no simple explanation for this advantage; the reality is that even poor urban children tend to enjoy meaningful advantages over their rural peers in a number of areas relevant to education. The much smaller gap in S2 is somewhat surprising, especially given the size of achievement differences in P3 and P6. But when one takes into account who is being compared in each location, the result is not surprising.' (Najjumba and Marshall 2013, p. 29).

Annexes A, B and E provide further details.

As an example of these details, one aspect of the validity of PLE is as a predictor of UCE results (since the PLE functions, in part, as a selection mechanism). In the data set we studied, there is a moderate relationship of PLE results and UCE results – that is, while high PLE performers tend to do well at UCE (and vice-versa) there are exceptions: high performing PLE students who do badly at UCE and low performing PLE students who do well at UCE. Such moderate levels of association are typical of those seen in other datasets where performance at one level of education is indicative of performance at the next. Our studies also included an initial look at these relationships using multi-level modelling.[8] These results suggest that there are some systematic differences in the UCE/PLE relationship across different UCE Examination Centres. We suggest that further exploration of these relationships, especially when it is possible to include those who complete the PLE but do not enrol for the UCE, will provide useful insights into regional differences in participation and attainment – differences that will then need to be understood before effective actions can be developed.

Such differences by region, rurality, socioeconomic status and so on are to be found in many parts of the world. For example:

- in MENA
  - 'In part as a reflection of the extensive inequality in the quality of primary and secondary education, the recent expansion in access to higher education in MENA developing countries has applied less to children in rural areas and those whose parents have low levels of education and are engaged in basic occupations than to children from urban and stronger socioeconomic backgrounds.' (Brixi, Lust and Woolcock, Trust, voice and incentives: learning from local success stories in service delivery in the Middle East and North Africa 2015, p 48)
- in France
  - 'At the same time, a concern for equality conceals the worst inequalities … a policy of having everything the same – learning programs, timetables, examinations … – that is

[8] See further details in Annex A and more technical information in section 2 of the appendix providing tables and graphs, including numeric results.

contradicted by daily realities, a sham policy that has no impact on social or geographical differences.' (Adapted from Toulemonde 2004).

From an assessment perspective, the issue is whether the assessment system only holds a mirror to these differences or if it, instead, contributes to them (and to what extent it does so). That is, the products of assessment may merely reflect the outcomes of inequities in the provision and practice of education, or the processes and products of assessment may both reinforce and contribute to inequities in outcomes.

Assessment may reinforce and contribute to inequities through a range of mechanisms, including:

- fostering competition (albeit more among students from the elite) rather than cooperative learning (K.-M. Cheng 2014); and
- assessment practices that privilege particular styles of learning, styles of expression and assumed background knowledge and experiences.

Testing and examinations are not neutral measuring devices. Many of their aspects have, and communicate, an underlying ideology: their content, the types of responses valued, the ways they expect test takers to behave (individuals working alone in silence), the assumptions they make about what test takers know about the background to questions, and the way that questions should be understood and answered. The family and cultural backgrounds of some students give them a head-start over others in terms of taken-for-granted familiarity and ease with the practices of tests and examinations.

The impact of assessment processes on inequity of outcomes can be seen in other countries:

- in Australia
  - 'The Competitive Academic Curriculum (CAC) has gained increasing influence over primary education… The form of the CAC has come to dominate areas of curriculum which used to have a different logic, for example music, art, health, 'design and technology'. The hegemony of the CAC is well shown by what happened in New South Wales with the designation of 'Key Learning Areas'; this became a vehicle for the creation of new academic subjects rather than (as I am sure was intended) a broadening of educational experience. The competitive academic curriculum became the subject of intense debate in the 1970s and early 1980s, when the 'new sociology of education' argued it was a key to the reproduction of social inequalities in education, especially the inequalities of class.' (Connell 1998, p. 89);
- In Indonesia
  - 'Indonesia's education assessment practices require further improvement if the system is to deliver the educational outcomes required by the emerging economy and the changing society, and offer fair educational opportunities. Without a broader assessment framework designed to underpin Indonesia's expressed educational goals, it will not be possible to adequately monitor progress in student learning, modify learning experiences and teaching practices where necessary, and evaluate the effectiveness of teaching at the classroom, school, district and national levels.' (OECD/Asian Development Bank 2015, p. 43).

4.5 Much of the knowledge and skills learners need today and for their futures has little or no role in formal tests and examinations.

There is currently a significant mismatch between what students actually learn, and the knowledge and skills valued and emphasised by the current assessment system, and what employers see as key skills or competencies for successful lives and careers now and in the future (Guillermo 2012).

For example, English is widely used in spoken and written forms in many fields of employment in Uganda today and in many people's daily lives. The skills in written English most privileged in the PLE examination (for example, the use of the past perfect tense in a conditional sentence – see the discussion in Annex A) are ones that are not in common use in many English speaking (and writing) environments in Uganda and elsewhere, are often not part of the English of many native speakers and are not part of modern business communications using English, whether nationally or internationally.

The knowledge and skills that learners need for successful lives and careers may change as Uganda's economy changes. The term '21st century knowledge and skills' is widely used to capture the idea of skills required for success in a modern, information and technology based economy, such as that envisaged in Uganda Vision 2040 (National Planning Authority 2007).

A Cisco, Intel and Microsoft project (Griffin 2012) developed the following framework for describing the skills learners need now and are likely to need more and more in the future.

- Ways of thinking:
  - creativity and innovation;
  - critical thinking, problem-solving and decision-making; and
  - learning to learn and metacognition.
- Ways of working:
  - communication; and
  - collaboration and teamwork.
- Tools for working:
  - information literacy; and
  - information and communication technology (ICT) literacy.
- Living in the world:
  - citizenship – global and local.

Annexes B, C and D (supported by reviews by NCDC: Kamwine (2012) and Kamwine (2015)) indicate that the sorts of knowledge and skills expected and valued by Ugandan employers, as well as the skills required in a more complex and modern economy, are largely absent in classroom and formal (PLE, UCE and NAPE) assessment.

The evidence set out in Annexes A, B, C and D shows that assessment, in practice, pays more attention to recall, whether of facts or of learned procedures, than to higher order conceptual thinking and understanding. The realities of practices, not policies or aspirations, mean that the assessment system does not deliver the emphasis on the higher order skills, thinking and understanding that the formal curriculum explicitly expects.

These are familiar issues elsewhere:

- in Vietnam
  - '…for the most part, because students are quite good at rote memorising, the information does stay in their heads, at least until the exams come. A bigger question … is whether students can deal with unstructured problems: can they think, analyse, and solve problems that aren't easy to describe and answer from information written on a board.' (Napier and Hoang, What we see, why we worry, why we hope: Vietnam going forward 2013, p. 41); and
- in countries participating in PISA
  - 'More than ever before, living and working in the 21st century requires the "four Cs" – creativity, critical thinking, communication and collaboration – but also digital literacy, proactivity, adaptability and open-mindedness. The question is whether schools today can foster these creative and social skills so that students are adequately equipped to meet these challenges. Many believe schools have changed too little since the 19th century to prepare students for the 21st.' (Echazarra, et al. 2016, p.7).

4.6 Ensuring integrity and authenticity in high-stakes assessments presents significant challenges

Every agency providing high-stakes assessments has to manage two distinct but related risks:

- loss of security of examination papers, cheating by candidates in any of many forms, and inappropriate behaviour by invigilators, markers and examiners; and
- loss of public confidence in the integrity and fairness of the results.

These risks are related but distinct because public confidence can be lost even if there are few actual incidents of loss of security or failures of integrity. And public confidence, once lost, is hard to regain.

Achieving public confidence in the fairness and integrity of classroom assessments (coursework or continuous assessment) used in whole or in part in high-stakes assessments requires an appropriate enabling environment (R. Allen, Australia - Developing the enabling context for school-based assessment in Queensland, Australia. 2012). It is common in many countries (but not all) to find little or no confidence in the fairness of the use of classroom assessment alone for high-stakes purposes.

Where classroom assessment is used in part for high-stakes assessment results there are usually procedures (scaling or external moderation) that are intended to remove any systematic bias and to give public confidence in the fairness of using these teacher-based assessments.

Pido (2004) provides a comprehensive report on the challenges of using coursework assessment in UCE subjects. These challenges include not only technical, procedural, data collection and training matters but also issues of low validity, limited reliability, lack of consistency, fairness in marking, integrity in results and problems with public perceptions.

Improving classroom assessment, especially as a replacement, in whole or in part, for summative assessment is a complex challenge. There are no simple solutions: official policy is not automatically reflected in implementation, nor is professional development sufficient.

The TVET sector seeks to develop competence-based curricula and successful continuous assessment while providing, through the work of the Uganda Business &Technical Examinations Board (UBTEB), the rigour and authenticity of assessment and quality assurance expected

internationally in this sector. UBTEB's reports clearly demonstrate recognition of the significant challenges – procedural, resource and capacity-building – involved in ensuring the authenticity (and quality) of class and project work, especially where this counts towards significant qualifications.

The challenges of managing risks to the integrity of assessment results are familiar in a wide range of countries:

- '...fraud in academic performance, and buying and selling academic marks/scores and academic degrees can occur, particularly in southern Europe and Russia… teachers play a crucial role in education … This role can be adversely affected due to corruption in managing teachers, which can take the form of bias, close relationships, friendship and bribery in appointment, rotation/transfer and promotion.' (Transparency International 2011, pp 6-7)
- 'Allegations of test-tampering and grade-changing by educators this year are on a pace to exceed the number of complaints made in 2014, continuing a rise in such allegations that began during the previous administration and has persisted .... The allegations come as New York City has scrambled to respond to a string of reports this year involving academic fraud and grade inflation, such as a high school that let students earn credits without receiving instruction and an elementary school principal who forged student answers on a state exam. Last month, the education department established a $5 million task force to closely monitor schools' test scores and how they assign credits.' (Wall 2015)

Each year, UNEB meets the challenges of minimising the integrity risks in the development, administration and marking of high-stakes external examinations across Uganda.

4.7 Many parents look for ways to ensure their children get the high scores needed to get into 'good' schools for the next stage of their education

Annex E summarises the main elements of public discussion (as reflected in the media and on the web) about primary and lower secondary assessment – its purposes, products and processes. A key focus is on the idea of a 'good' school as one where students get high scores in PLE (for entry to a 'good' secondary school) and high scores in UCE, for entry to 'A' level. For example:

- 'Following the release of Uganda Certificate of Education (UCE) results last Friday, the next task for students and parents is to find a school with a record of performing well at the next level; Uganda Advanced Certificate of Education (UACE), the gateway to university and tertiary institutions. Today, increasingly thoughtful employers require personnel that have relevant job skills and ability to improve the productivity of their companies and enterprises regardless of their academic performance. Nevertheless, academic excellence remains an integral yardstick for possible employment in many institutions. As such, every student and importantly parents, wish and are always out to take their children to the best performing schools academically.' (http://www.monitor.co.ug/News/Education/Top-100-A-Level-schools-over-the-last-four-years/-/688336/3056764/-/q7bo8z/-/index.html)

There are obvious links between the capacity of parents to meet these aspirations and the equity issues discussed earlier.

Some Ugandan parents told us how they wanted their children to have a fully rounded education but at the same time recognised the importance for the young person's immediate future of giving priority to coaching to pass high-stakes examinations.

These are familiar issues across the world (although not everywhere – Finland appears to be somewhat of an exception). For example:

- in Shanghai
  - 'Parents still look for ways of ensuring their children get high scores, and teachers may very much gauge their teaching in response to expectations about public examinations and university entrance.' (K.-M. Cheng 2014, pp 3-4)
- in Singapore
  - 'The majority of parents [in Singapore] are spending hundreds of dollars every month on private tuition, despite knowing that extra classes may not significantly help raise their children's grades. And even before the children attend formal school, many children under seven are spending hours in extra classes every week brushing up on their English and maths … School examinations such as the O levels and PSLE could also be adding to the pressure for parents to ensure their children do well in them… That pushes parents to pay for expensive private tuition…' (Davie 2015)

5 A 'systems approach': What is it and why is it necessary?

Uganda's assessment system is a complex and interconnected web of agencies, people, relationships and settings, in which assessment has many forms, many purposes and many functions.

While no-one set out to design the totality of the current assessment system, it functions as if it had been carefully designed to achieve exactly the results it achieves now – results that are not aligned with Uganda's needs now and in the future.

Within the assessment system there are complicated chains of causes and effects. Change is not, therefore, simply a matter of changing a single element (agency or agency action).

We propose a continuous improvement approach, grounded in a focus on outcomes and stakeholder engagement.

The specific issues underpinning each of the seven headlined descriptions in this report should not be understood in isolation. They form aspects of a whole – Uganda's assessment system, which reflects many aspects of Uganda's history, context and culture.

By themselves, individual changes to elements of assessment/examinations, importing ideas and processes from elsewhere, or replacing one kind of assessment with another will not lead to an adaptive, resilient self-improving system, working within its culture, context and resources. It is important:

'not to do what analysts so often do: focus on a few key facets of the system in a search for magic bullets, ignoring the fact that the success of these particular factors is made possible only by the myriad other features of the system that gives it its particular gestalt. We cannot, they say, really understand how the Shanghainese built such an effective education system unless we understand it as a system that is more than the sum of its parts. People educated in the analytical methods of the sciences typically analyze everything, decomposing systems into their constituent parts and try to estimate the contribution of each to the effect on student achievement. By all means, do that, say Cheng and Jensen, but, unless you grasp holistically the way the whole comes together, unless you grasp the motivating spirit of the system, you do not really understand anything very important.' (Tucker, Chinese Lessons: Shanghai's Rise to the Top of the PISA League Tables 2014, p.3)

5.1 A conceptual description of Uganda's assessment system – a complex and interconnected web of agencies, people, relationships and settings

The approach developed in this road-map is a systems approach, which seeks to understand what happens as a complex and interconnected web. This is summarised in the following diagram, of the assessment system as a complex web of government and non-government agencies and people, relationships and settings that sits within the foundational context of legislation, policies, priorities and resources set by government.

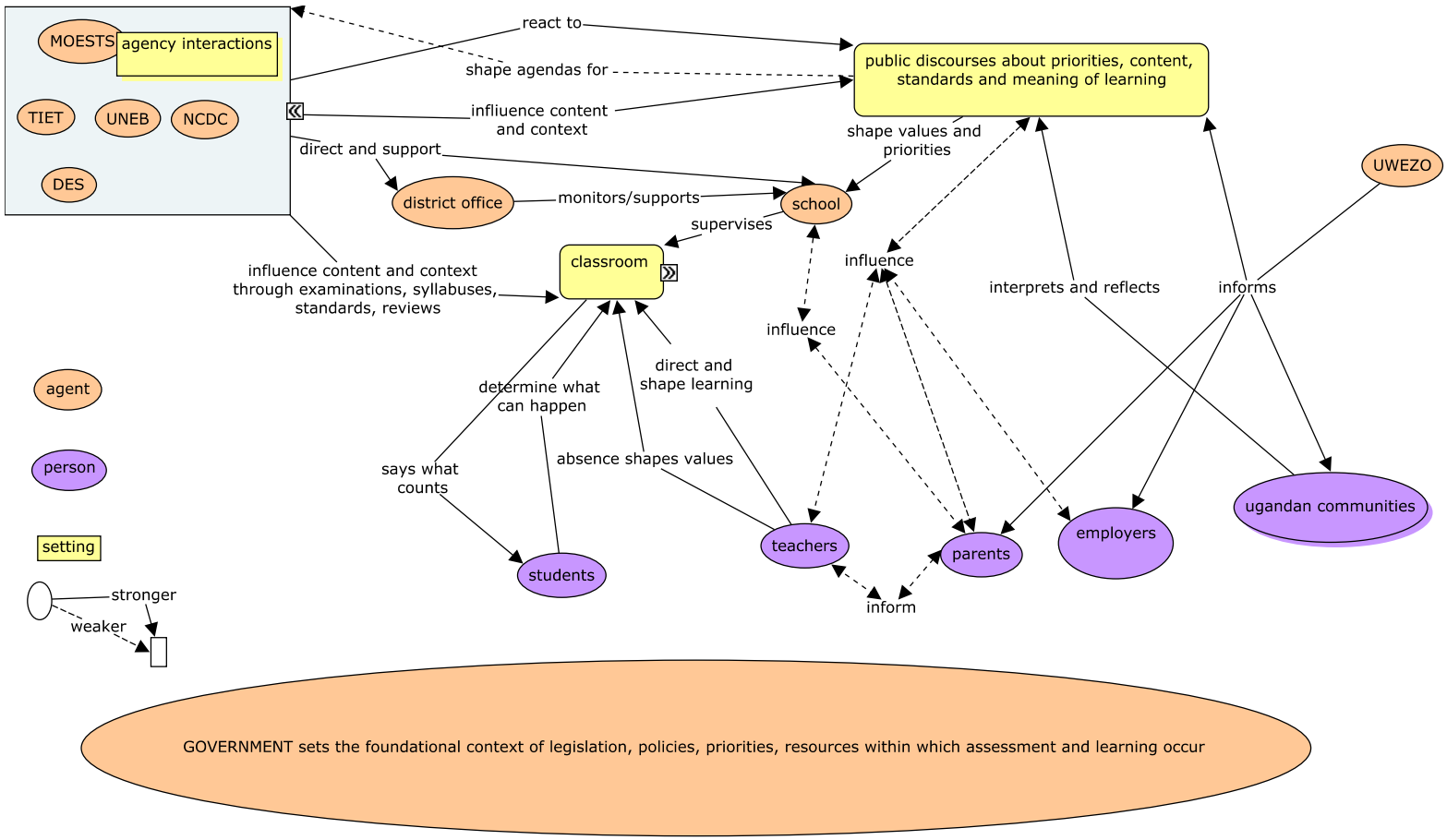
The diagram is complex enough, but it is, of course, a simplification of an even more complex reality. However, such a diagram is a useful tool in developing a systems approach to enhancing Uganda's assessment system. First, it indicates the major influences on each of the three settings, and how these influences connect – which is useful when planning developments in ways that take account of current realities. For example, it is clear from the diagram that we should not expect classroom practices to change simply because there is a change in the official curriculum – there

are many other factors at work: for example, a school's, and its teachers', values and priorities are shaped by parents and by the public discourse. Second, the diagram suggests that some connections are weaker than others, and some that might be expected to be there are absent. For example, the diagram reflects our understanding of the limited influence of the ideas, values and priorities of employers regarding the content and standards of education in Uganda. As a further example, the diagram reflects that we have not seen strong indications that Uwezo's work influences teachers effectively.

The diagram can be used as a tool to suggest ways in which support for changes can be developed and broadened – what Andrews, Pritchett and Woolcock (2016) describe as '*building broad agency solutions in the process of identifying problems and finding and fitting contextually appropriate solutions*.'

The purpose of this section of the report is primarily to be descriptive (rather than evaluative, commenting on strengths and weaknesses) of the main elements of the assessment system *as a complex system* and those key results that should be considered in developing a road-map for enhancement.

Figure 1: A concept map of Uganda's assessment system



5.2 Different impacts in different settings

The direct impacts of the assessment system – its immediate results – appear through what happens in three principal settings:

- classrooms;
- public discussion; and
- interactions between and among agencies.

First, the significant result of what happens in the *classroom* setting is *what students actually learn*, which is directly influenced by assessment of, as and for learning, assessment that may be formal or informal. Secondly, the significant result of what happens in the *public discussion* setting is *the shaping of the political environment* within which agencies – including schools – work. Lastly, the significant result of what happens in the *agency interaction* setting is the *balance across agencies of power and influence* over roles, responsibilities and resources. These immediate results in turn influence the rest of the assessment system.

5.3 Roles within the assessment system – agencies and individuals

There are three categories of agencies with major roles in Uganda's assessment system: government agencies with a Uganda-wide remit, non-government agencies with a Uganda-wide remit and agencies with a local remit (such as schools and district offices).

The actions of agencies include:

- the administration of examinations, tests and other assessments and the release of results (individually and in aggregate form);
- the development and publication of syllabuses;
- the development and application of standards for the review of the performance of schools and teachers; and
- the coordination and supervision of teachers in each school.

The actions of individuals include:

- students' actions in and responses to classrooms, schools and learning;
- teachers' management of classrooms;
- teachers' absences from classrooms;
- parents' decisions about choice of schools;
- the response of members of the community to public discussion about results and schools; and
- the views employers express about the skills of students entering the workforce.

5.4 Realities and aspirations

While no-one set out to design the totality of the current assessment system, it functions as if it had been carefully designed to achieve exactly the results it achieves now. The design and implementation of change requires understanding current realities – why and how they interact to produce the results they do. This does not mean that these results align with the *aspirations* of each agency and individual. Achieving such aspirations can be limited by current realities. For example:

- a parent may want a child to have a fully rounded education but also recognises the importance of succeeding in the system as it currently is;
- a curriculum development agency may want the assessment system to drive comprehensive implementation of the official curriculum in all schools for all students but also understands the impact of examinations on what is taught and learned;
- an examination agency may want to make significant changes in the content and format of examinations but also recognises how its opportunities are limited by external constraints, community expectations and resources; and
- Government may want an assessment system that works satisfactorily at a much lower recurrent cost but also recognises the constraints of public expectations.

5.5 The key results and purposes of the current assessment system

The results actually achieved by the current assessment system allow us to infer its *actual* purposes. Its actual purposes are not the purpose or purposes that are officially stated. For example, one of the key actual purposes of the current assessment system is to produce public comparisons of schools. Regardless of intentions or policies, the current assessment system in fact produces information that is used to compare schools.

The seven headlined descriptions discussed above describe central features of the current assessment system. The key results of the assessment system, which reflect these features and their interaction in the ways shown in the diagram given earlier, include the following:

- too many students who are **successful in school** do not learn enough of the skills and knowledge aligned with the current and future needs of employment and the further development of Uganda's economy;
- public discourse about education focuses more on scores than on what is learned and its alignment with the needs of individual and community success now and in the future; and
- reports by agencies often do not lead to effective action – at central, district and local level.

These results are a consequence of the assessment system as it is, not a consequence of a failure or fault of any individual agency or person.

We have selected these key results for attention because the focus of this report is on *enhancement.* There are other key results not listed here, of course, ones with immediate significance to students, parents, schools, the community and the government.

It is important to emphasise that the first key result listed above focuses on those who succeed, not the many who fail. One effect of the current assessment system is that it drives the learning of those who **succeed** in school in directions ***away*** from the learning they need in terms of the skills requirements of current and future employment. Consultations with employers during curriculum reviews have identified the importance to employers of students' having 'generic skills', and their perception that students who have **succeeded** in education up to Secondary 4 (S4) level have **lower** levels of these skills than new recruits in general (Guillermo 2012).

In our view, this reflects ways in which the many features of Uganda's current assessment system work together to drive instruction and learning, defining and rewarding success and, for example, marginalising vocational learning. As some have described it to us, the top students study physics, chemistry and mathematics – learning knowledge and skills (often quickly forgotten after the exam) that do not connect well, or clearly, with work, the needs of employment and the skills required of a successful entrepreneur. The consequent impact of the assessment system on those who are not

successful is that many have little positive achievement and they have not learned from their school experience the skills and knowledge they need for success in life and career.

These three key results are not the results that Uganda needs its assessment system to produce in the 21st century.

First, skills and knowledge aligned with the current and future needs of employment and the further development of Uganda's economy include those required in agriculture and small- to medium-sized enterprises in the non-farming sector, emergent and growing areas of skills demand, including the hospitality industry, the ICT sector, the business management and financial sector, mining and engineering, oil and gas, and environmental technologies. Such skills and knowledge include not only practical competencies but also soft skills, such as communication, computer literacy, customer care, problem-solving, work attitudes and ethics (Ministry of Education and Sports 2011). Uganda's Vision 2040 (National Planning Authority 2007) sets out an expectation that Uganda's education curriculum, examination and instruction methods will be revised, and underlines the need to be responsive to market demands.

It appears clear to us that the need for change has been recognised – what is necessary is a **process** that is able to achieve these changes in a **timely, effective and affordable** way.

Secondly, schools, districts and teachers respond to and reflect the community's discourse about education (they are all part of the community). A focus on seeing the success of a school in terms of scores, rather than in terms of learning – especially a focus on 'top scores' rather than on learning for all – reduces the capacity of schools and teachers to achieve the expectations set out in curriculum documents. The road-map in this report sets out a process whereby over time the public discourse can be more usefully aligned with what is important in education.

Thirdly, we found that Uganda's assessment system in 2016 has a focus on examinations designed for selection and we identified consequent negative impacts on teaching and learning. The Education Policy Review Commission chaired by Professor Senteza Kajubi made similar observations in 1989. The underlying stability of the current assessment system continues despite various reports identifying problems and proposing technical solutions, and despite well-written policy and curriculum documents. There is little value in having reports with apparently sensible (albeit narrowly technically based) recommendations that are not turned into effective action. Indeed, this can have negative consequences in itself: when the latest, freshest report says (as if it has been discovered for the first time) something that those with long experience know has been said many times without leading to action. As well, reports about changes and capacity developments too often leave out, or make only passing and superficial reference to (for example, 'subject to suitable resourcing'), the broader social and historical context in which agencies, such as assessment and examination agencies, operate.

5.6 Achieving change

Since the assessment system is a complex web of interactions and relationships, there are complex chains of causes and effects. Change is not, therefore, simply a matter of changing a single element (agency or agency action) in the system. As Ben Jensen has pointed out, when discussing Shanghai's success:

'It is certainly true that Western academics try to identify discrete policies and practices, measure their independent effects and then assert that one or another of them is the answer. But anyone who actually works in any sort of large system or organization realizes that it is much more complicated than that. I've seen a lot of systems—and Australia is very good at this—that think one

or two particular programs constitute our silver bullet. Whereas when you look at Shanghai, you see there is a very comprehensive, coherent long-term strategy embracing many elements, all of it clearly focused on teaching and learning.' (Tucker, Chinese Lessons: Shanghai's Rise to the Top of the PISA League Tables 2014, p.27)

There is no single action or set of technical changes borrowed from elsewhere that will produce the assessment system that will meet Uganda's needs in the 21st century. For example, it is clear that UNEB well understands the importance of including items in its examinations that assess higher order thinking. In the context of what currently actually happens in many classrooms (over several years) – widespread and frequent use of practice examinations and tests that do not assess higher order thinking – students (and their teachers and parents) would not feel fairly treated if there were many such items in UNEB's examinations. The influence of past papers on teaching, learning and the next set of examinations is a well-known phenomenon. Change, which can occur (see Hong Kong for an example), requires a systems approach – one that takes into account the existing attitudes, values, expectations and practices of teachers, students, schools, parents and agencies, and that builds the required changes in attitudes, values and practices – not a task that is accomplished with a simple policy change or a programme of technical capacity-building.

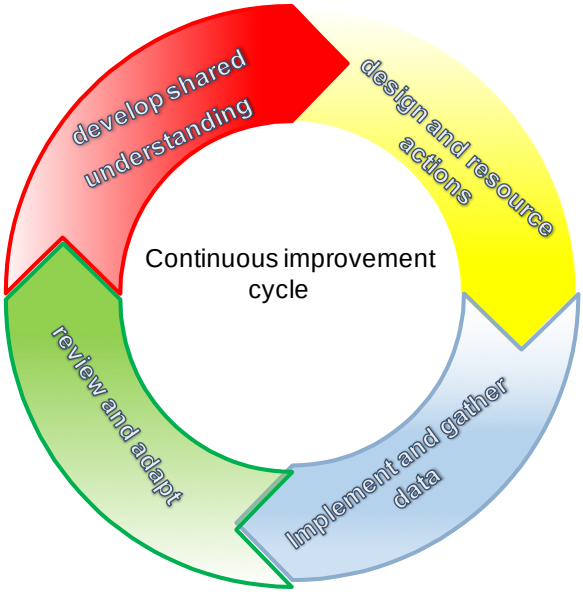
5.7 The idea of continuous improvement

The road-map provided in this report sets out a continuous improvement process, one that is monitored against standards, with independent reviews of effectiveness. Such a process will lead to a continuing accountability to provide clear explanations of why something stakeholders have agreed is important has not happened as it should, and how this situation will improve. Continuous improvement is the idea of improving not through turning everything upside down but through a continuing process of analysing and understanding, planning, acting and evaluating[9]. This is a cyclical process, one that goes on, not one where a single set of changes is designed and implemented in a linear fashion. Continuous improvement is based on the assumption that we never reach perfection (partly because the world keeps changing around us) but we can keep on developing and implementing improvements. As we use the term in this report, continuous improvement can, over time, lead to major changes.

The following diagram emphasises the iterative, sustained and adaptive aspects of this approach.

[9] Such a process *strengthens* the capacity of an agency to meet its legislated mandate, since it provides for agreement across stakeholders about the nature, importance and resourcing of improvements.

Figure 2: Continuous improvement cycle



Key elements of our continuous improvement approach, described in more detail later, are

- developing shared understanding across stakeholders about
    - what the various elements of the assessment system are expected to achieve (using a standards approach to achieve a focus on outcomes and outputs, rather than on processes – a focus on results rather than technicalities)
    - the most important gaps between what happens and what it is agreed ought to happen;
- working with stakeholders to develop and agree on the nature and resourcing of interventions to reduce these gaps;
- implementation of agreed and resourced interventions, together with gathering data about effectiveness and impact from the start of each intervention; and
- reviewing the effectiveness of interventions – and adapting them to fit a better understanding of the underlying issues and changing circumstances.

6 A future assessment system – the 'destination' of the proposed road-map for enhancement of the current system

We propose a multi-step process, based around explicit standards, independent reviews, incentives and stakeholder engagement that will make Uganda's assessment system robust, resilient, self-sustaining and self-improving.

6.1 A robust, resilient, adaptive, self-sustaining and self-improving assessment system based on agreed standards

An assessment system that will be able to produce the results Uganda needs and wants in the future under changing economic, social, political and external circumstance is a *robust, resilient*, *self-sustaining* and *self-improving* assessment system, grounded in Uganda's culture, history and context.

Such a system is characterised by:

- clear, explicit and published standards for the processes and products of the agencies with responsibilities within the assessment system;
- open, honest and transparent reporting against these standards;
- a climate in which reporting against standards is used for continuous improvement rather than blame or sanctions; and
- positive and negative incentives institutionalising and reinforcing the appropriate practices of individuals and agencies.

These features make this an *adaptive* system, one that will accommodate anticipated and unanticipated external changes in the economy and the society – a system that will identify and respond to unintended consequences of designed interventions.

As is clear from the set of actions described below, the proposed destination is an *evolution* of the current system, not a revolution. Radical, revolutionary change in education is very disruptive, normally very expensive and often has unintended and unexpected consequences – usually negative.

Having such a system will develop actions on the basis of consensus about the detailed developments needed to change the key results of the current assessment system and the nature and significance of the seven headline issues described above.

We understand from a draft concept note entitled 'Proposing Principles for Performance Conditionality, Accountability and LG Support within the Consolidated Grant Framework' that Uganda is developing a performance management process in a major area of government activity that has some key elements in common with the ones listed here, including a shared focus on outcomes, continuous improvement and built-in third party validation.

6.2 Standards for the processes and products of the agencies with responsibilities within the assessment system

Standards are statements that describe the *intended* qualities of the *most important* processes and products of an agency. Standards should be brief, clear and direct. Setting such standards does not mean that they will be achieved fully and completely – it means that there is a set of clear statement about what is wanted, that there is a public expectation of continuous improvement efforts directed towards these standards, and there are good, transparent, accounts of what happens and what the factors are that explain the gap between what actually happens and the standards. Such factors will include factors that are *within* the agency's direct control and influence, and factors that are *outside* the agency's direct control and influence – such as community attitudes and values, external events and externally determined resource allocations.

It is important that the statements of standards do not have built into them elements that make them ineffective for continuous improvement purposes. Such elements include the use of terms such as 'satisfactory', 'improving', 'providing students with opportunities', 'maximising', 'strategic', 'should'. For example, the statement that 'a school should provide its students with the best possible opportunities for education' does not establish an adequate basis for reporting, continuous improvement and review.

Some illustrative examples of standards at various levels follow:

- the agency's decision-making processes are transparent and accountable;
- the agency has a documented and reviewed quality system that incorporates an evidence-based continuous improvement process;
- the examination agency's assessment methods provide a valid and authentic measure of the required knowledge, skills and understanding, giving students opportunity to demonstrate their capacity to meet the full range of requirements;
- curriculum documents reflect community and employer expectations of high quality courses; support student participation and achievement; and are useful to, and used by, practitioners;
- the school knows and acts on information about the extent to which, each year, it produces students who
  - *are well connected with school and education*[10]
  - *have made at least the minimum acceptable progress required by the curriculum*
  - *are ready for the next stage in their education*; and
- the Ministry knows and acts on information about the extent to which, by the end of primary school, all young Ugandans can read and write in English, use basic everyday mathematics skills and everyday technologies, and are ready to succeed in secondary school.

  The development of statements of standards is a matter for collaboration among the full range of stakeholders in the current Ugandan assessment system. Developing such statements will contribute to building a shared understanding of the current situation and agreement about the need for change and its basic directions.

  Agreeing on standards, particularly those with a focus on outcomes (the external impact of what an agency does), together with information about current realities, will help to build a

[10] **School connectedness** is 'the belief held by students that adults and peers in the school care about their learning and about them as individuals'. There is a clear evidence-based association between school connectedness and both health and school attainment.

shared understanding of the most important gaps between what does happen and what ought to happen. Such important gaps are likely to be difficult problems that cannot be solved by policy statements or by providing resources to achieve technical developments (although technical improvements will contribute.).

7 Strengths in the current system that can be used as a foundation for the proposed approach

As well as the evident wide and successful experience and skills in delivering assessments in Uganda's complex and changing environment, there are particular significant strengths within the current system that can be built on in the development of the standards-based approach proposed. These include:

- the standards approach for reviewing the work of schools and teachers developed by DES (see, for example, the DES Annual Report for 2012 and the clearly expressed statements of standards for schools and teachers);
- the reviews of reliability, validity and integrity carried out by UNEB (the examination agency);
- the reviews of syllabus implementation carried out by NCDC (the curriculum development agency); and
- the community-based evaluations provided by Uwezo (a citizen-based movement with a focus on literacy and numeracy).

These are foundational strengths because they demonstrate existing capacity for:

- developing and applying standards for process, outputs and outcomes;
- reporting the quality of high-stakes assessment practices;
- evaluating implementation of curriculum in terms of gaps between intentions and practices; and
- engaging the community as active stakeholders.

8 Elements along the road

Within the idea of continuous improvement, we identify a set of actions that are influenced by and build on existing capacities and shared understanding amongst stakeholders.

In this approach, technical developments and capacity-building occur not as a starting point but as a *consequence* of an agency's reporting against a set of standards for its processes and products.

The following elements are listed in a specific order, with the later elements building on developments in the earlier ones. Consistent with the basic principle of continuous improvement based on review and monitoring, these elements do not constitute a set of fixed steps but are, rather, a set of actions that are influenced by and build on the capacities, understanding and agreement among stakeholders developed by earlier steps. This is a process that starts now and continues over the next five to ten years.

The initial elements are:

1. further data gathering and analyses to develop an understanding of the key issues and factors influencing the present situation (these include further classroom observations and technical analyses of examination data);

2. collaborative work with agencies to develop sensitisation to the idea, management and challenges of continuous improvement approaches; and

3. collaborative development across stakeholders (including employers and the community, as well as central and local agencies) of standards for the processes and products of major agencies.

Once some standards for one or more major agencies have been developed and provisionally agreed, it is then possible to conduct

4. external reviews of the extent to which these standards meet requirements for clarity and accountability; and

5. capacity-building to support the reporting by agencies against these standards, including the development of proxy measures of quality elements of the standards (for example, progress measures for schools).

Reports against these standards can then be made. A report of the work of an agency against a set of standards has two key aspects:

1) a self-assessment by the agency, identifying

- quantitative and qualitative data showing the extent of the match of the agency's processes and products against these standards
- evidence showing the impact on the extent of this match of the work of the agency (factors inside the agency's control and influence)
- evidence showing the impact on the extent of this match of factors outside the control or influence of the agency; and

2) an independent external review of the report to identify:

- commendations – features shown in the self-assessment report that meet the specifications set by the standards
- affirmations – directions for improvement identified in the report
- recommendations – suggestions for further developments that may be considered.

The process of self-assessment and external review against collaboratively agreed standards has direct benefits to the agencies involved. First, the standards for an agency are understood, agreed to and supported by external stakeholders, not internally defined. This helps support fairness and transparency in the allocation of scarce resources. Secondly, self-assessment followed by an external review provides an incentive to an agency to set out clearly, openly and honestly what it knows about how well it does its work and how it can feasibly improve. Self-identification of problems together with opportunities for improvement become an occasion for the external review to make positive comments about the work of the agency, rather than finding fault or laying blame. Thirdly, the process leads to an incentive for the agency to improve, to become more efficient and effective, and to show that this has occurred. This has a positive impact on the lives and careers of those working in the agency.

Capacity-building to support agencies in the principles and processes of self-assessment will be required, and capacity-building in the principles and processes of independent external reviews will also be required.

All these reports will be published documents. Capacity-building for key priorities for technical improvements identified in these reports will be required.

One or two years after the publication of these reports it will be necessary to conduct and publish external reviews of the extent to which the opportunities for improvement identified in reports have been implemented and the internal and external factors affecting implementation.

Starting at more-or-less the same time, it will be possible to work with government to enhance understanding of possible approaches to transparency and incentives leading to collaborative development of recommendations to government of a system of positive and negative incentives for agencies to report honestly and act on opportunities for improvement.

A review of the allocation of roles and responsibilities among central government agencies (for example, the separation of curriculum and assessment) should be conducted at some stage during the implementation of this road-map.

In discussing assessment, technical aspects of examinations, standardised tests and classroom assessment are often thought of first and programmes of change are developed around designing and implementing new, different or improved techniques of assessment. Such an approach leads to first-order interventions on single aspects of the complex, interdependent web described above, usually with unintended and unanticipated results. As a simple example, a change to the means of high-stakes assessment (for example, the use of continuous assessment) will have an impact on the integrity and other aspects of quality assurance of results, an impact that is not mitigated by training and exhortation but will, fairly or not, have an influence on public perceptions of the value and meaning of the products of assessment and, in turn, on the outcomes of the assessment system as a whole.

In the approach outlined in this road-map, technical developments occur not as a result of external experts providing authoritative direction but as a *consequence* of an agency's identifying opportunities for improvement of their processes and products against a set of standards that have been developed collaboratively and agreed by government, with later reports showing the extent to

which such technical developments have in fact led to the expected improvement. For example, we were able to analyse patterns of student responses within an examination by using a limited sample of 300 marked scripts in each of four PLE papers (see Annex A). UNEB successfully administers a large-scale examination system each year – meeting massive logistical challenges in a limited timeframe and with limited resources. The marks on individual questions are not included in the UNEB databases – and so the techniques of modern psychometrics cannot be applied on a regular and routine basis. These techniques could be applied to a sample of marked scripts in each subject, at a modest cost each year. The results of doing this would only be of sufficient real value if there were a clear statement of the purposes of the examinations (the outcomes that putting resources into having these examinations is expected to have) and if the psychometric techniques were then used not as an end in themselves but as a means to ensuring that the examinations better achieve these purposes and outcomes.

9 How the road-map responds to the current assessment system's headline features and key results

The key problems are difficult, complex and multi-faceted, with no perfect solutions: so-called 'thick' problems.

Consistent with our systems approach and with ideas of continuous improvement and stakeholder engagement, we illustrate ways in which developments might occur and be improved iteratively.

The road-map provides a process of continuous improvement, one that is iterative and adaptive, and that is a continuing cycle of understanding current realities, designing interventions, and implementing, reviewing and adapting those interventions.

The collaborative development of explicit standards for the processes and products of agencies will develop agreement amongst stakeholders about the most important outcomes expected of both central agencies and of schools. This is the foundation on which to build shared agreement amongst stakeholders about what currently happens, why it happens that way and the extent to which it meets Uganda's current and future needs.

Reviews against such standards can then lead to agreement across all stakeholders about:

- the size and importance of gaps between the expectations of the standards and what actually happens; and
- how the actions, processes and products of central agencies and schools contribute to the features and results of the Uganda assessment system.

In turn, this can lead to agreement about which gaps between expectations and realities are large enough to require effective actions (including resourcing), and agreement about current actions and processes that are not sufficiently cost-effective to justify their continuation.

The road-map responds to the headline features and key results described above by identifying the steps in building capacity for a system in which Uganda's stakeholders can develop a shared and generally agreed understanding of the nature, importance and underlying causes of features of Uganda's assessment system requiring significant changes, and the design, implementation and review of appropriate interventions.

The capacity to develop a shared, agreed and nuanced understanding of the nature and causes of the sorts of problems identified in our headlines and key results[11] is the foundation on which priorities for effective actions can be agreed, implemented, reviewed and adapted.

Most of these problems are what are sometimes called 'wicked problems': so-called because they are complex, multi-faceted and difficult to solve, and because an effort to solve one aspect of the problem may lead to other, unexpected, difficulties. For example, changing the current teacher practices in assessment we describe in Annex D will not be achieved by simply issuing instructions or by organising a widespread programme of teacher professional development. What teachers currently do in assessment in Ugandan classrooms reflects not only their individual abilities and capacities but also, predominantly the complex set of multiple interacting factors and influences we have sought to describe. Effective action must be built on a clear understanding of the current situation, agreement amongst stakeholders about the importance and value of changes, and the development of first-order interventions that can be resourced and will be cost-effective.

[11] Sometimes this idea is called 'owning' the problem. Communities that 'own' their problems are better placed to develop useful responses to them.

This is why this road-map recommends:

- an adaptive approach;
- a collaborative, consensus-building approach;
- a continuous improvement approach;
- a standards-based approach; and
- an approach based on publicly reported reviews.

In summary, the steps involved in this change process are:

1. agreement amongst stakeholders that a particular feature is really important and worthwhile improving or changing (i.e. it is a 'problem');
2. using data from reviews of agencies' processes and products to develop a shared understanding of the many reasons for, and causes of, this feature and its interaction with other aspects of the assessment system;
3. using this shared understanding to agree on some feasible and affordable interventions that will have a significant effect;
4. implementing these interventions;
5. reviewing the impact of these interventions; and
6. adaptation, adjustment and refinement of the interventions (or identification of a deeper 'problem').

For example, aligning assessment practices at all levels (examinations, classroom assessment, practice tests) with the ideas and ideals of 21st century skills (creativity and innovation, critical thinking and problem-solving, communication and collaboration) requires agreement across all stakeholders (including schools and teachers, as well as the broader community including employers) about:

- the importance of, for example, creativity, problem-solving and collaboration, relative to other current in-practice priorities;
- the costs (resource, change-management, opportunity) involved;
- who will bear these costs; and
- what useful features of present practices will be given less (or no) attention.

The complexity of making such changes in assessment practices is reflected in the well-known impact of past examination papers on teacher, student and school practices. The examination paper in one year creates an expectation about the types of question to be expected the next year[12]. Adding new types of question creates (usually with some difficulties) an expectation about the following year. The question in one year's examination testing a capacity to work on unexpected, unfamiliar, unrehearsed problems[13] becomes, in later years, a question testing the capacity to give the 'right' response using a well-rehearsed approach – something that leads to coaching and memorisation.

A further challenge arises from the interlocked nature of the assessment system: should such changes take place from the upper secondary downwards or from the primary level upwards?

[12] In many parts of the world – including, we believe, Uganda – teachers with a reputation for correctly anticipating the types of questions in each year's examination paper are highly regarded by schools, colleagues, parents and teachers.
[13] For example, applying knowledge to an unfamiliar situation.

What happens in primary is, in part, a preparation for learning in the secondary level, and onwards to higher education. Substantial changes to primary education could leave students less well-prepared for the realities of later stages of education, and could lead to pressure on primary schools to make only token moves in the direction of the desired changes. Changes at secondary level may not be easy for students who have successfully negotiated primary. However, starting at the upper secondary level may be easier and will lead to demands for change at the lower secondary and primary levels. Changes in curriculum, and hence in examinations, at the lower secondary level may lead to demands for changes in assessment at upper secondary and primary levels.

Demand-driven change processes are usually much more effective than supply-led processes, although education has a long record of preferring the latter. Changes in practices that teachers and schools see the need for and demand are much more likely to happen effectively.

Changing examinations and other forms of assessment to include assessment of unrehearsed problems[14], higher level concepts, creativity and collaboration will not happen solely with a change in the techniques and technicalities of assessment. It can, however, happen over time, through the agreement and active contribution and engagement of all stakeholders (including parents, teachers and the community), through well-designed interventions or changes, through review of their effectiveness and through further adaptation.

Designing interventions or changes will be based on a clear understanding of why schools use the practice tests they currently use and how these meet needs of schools, teachers and parents.

Given such an understanding and given appropriate resourcing, a set of interventions might consist of a planned programme coordinating incremental changes over time in the content and format of examinations, including:

- the provision of appropriate practice tests across grade levels at a price competitive with the price of commercial tests;
- the professional development of teacher leaders at the district level;
- the fostering of the active engagement in the change process of community stakeholders, such as parents and employers;[15]
- adapting the programme of changes in response to systematic information about its effects – intended and unintended.

As a further illustrative example, an enhancement process might take the following form:

After review of the information in this report, stakeholders may agree that the PLE is an examination that:

- focuses on identifying the highest performing students in terms of a set of skills that are not aligned with the requirements of current and future employment and Uganda's economic development;
- does not provide a basis for thorough and reliable information about a school's, or a student's, performance until students have completed seven years of school; and
- involves students in over nine hours of examinations.

[14] For example, questions that require students to apply knowledge to an unfamiliar situation or questions requiring students to develop a process that they have not previously practised.
[15] In Uganda, most schools were founded by religious organisations. These are key stakeholders.

This is a complex, 'thick', problem. The development of an iterative, adaptive approach to this problem could start with consideration of the option of abolishing the PLE and replacing it with classroom-based assessment for formative and summative purposes. Stakeholders might then identify that this option would be likely to:

- substantially reduce the expenditure on examinations;
- remove the provision of any reliable information about student learning outcomes at a student or school level until the UCE level (given the current features of the classroom learning environments);
- remove the current possibility of developing progress measures for secondary schools;
- prevent the development of learning outcome measures to support performance management systems at school and district level;
- produce uncertainty amongst parents about the outcome of primary schooling; and
- remove a major driver of what currently happens in classrooms from P4 to P7 without an effective replacement for it.

Stakeholders might then develop another option. This could be based on the idea that, clearly, the resources (money, student learning time, teacher time) required by the PLE could be redirected at the same or slightly lower cost into high quality valid and reliable assessment at, say, the P4 and P7 level, using at each level four formal assessments, each being about one hour in length, for a total of eight (rather than nine) hours of assessment: four hours at P4 and four at P7.

Under this scenario, stakeholders would then agree on:

- the quality standards expected of these assessments in terms of their reliable and valid measurement of major and important learning outcomes (competences/core skills). (Published reviews of the performance of the examinations agency would then include reviews against these quality standards);
- how the results of these assessments would be used in published reviews of school performance (including progress measures);
- how to manage the transition issues, given that students in Grades 4 to 7 have been preparing for the PLE in its present form; and
- how to manage the implications for assessment at the UCE and UCEA level.

Given the value to parents of reporting in terms of scores, the prevalence of commercial tests of varying quality and the realities of many classrooms (numbers of students, varying levels of skills of teachers), stakeholders might agree on producing practice tests (aligned with the curriculum and the new tests) of a high quality at a price that makes them competitive with commercial tests.

This process outlined here responds primarily to the following features and key results:

- *in practice, little priority is given in classrooms to matters that are not closely aligned to the demands of examinations;*
- *teacher absences from classrooms are a fundamental constraint on classroom learning and assessment processes;*
- *much of the knowledge and skills learners need today and for their futures has little or no place in formal tests and examinations; and*

- *too many students who are **successful in school** do not learn sufficient skills and knowledge aligned with the current and future needs of employment and the further development of Uganda's economy.*

10 Implementation of these proposals

Successful implementation is going to require not only resourcing but, equally importantly, governance, support, risk management processes and a focus on outputs and outcomes.

The effectiveness of a high -level steering group will be critical to success.

Some significant risks can only be mitigated with significant Government leadership.

Implementation will require continuing appropriate funding and technical expertise in the required capacity-building. Experience suggests that the major risks to successful implementation are that:

a) standards will be written with built-in excuses (for example, 'students will have opportunities');

b) no (effective) actions will be taken;

c) there will be a greater emphasis on the relatively easy, 'thin' problems (for example, providing technical capacity-building support) rather than on developing actions that will provide some effective approaches to the 'thick', difficult problems;

d) reports will be seen as criticism of, or as assigning blame to, agencies or individuals, whether fairly or unfairly;

e) reports will be sanitised before publication or not published at all; and

f) the management and funding of the implementation of the road-map will not be sustained and strategic.

Elements in the road-map are intended to mitigate the first three risks (a, b and c).

Mitigation of the other risks (c, d, and e) will require significant leadership from the Ugandan Government.

Mitigation of the risk that the management of the implementation will not be sustained and strategic may be achieved by establishing a continuing coordinating group, with representatives from major agencies and funders, to provide oversight and to report to the Ugandan Government and the wider public on its work.

It will be important to establish a suitable governance and management structure for the implementation of this road-map that takes into account:

- its significance to the attainment of the Ugandan Government's priorities for education;
- its timeframe of five or more years;
- the wide range of key stakeholders, including government and non-government agencies (central and local), districts, schools, employers, teachers, parents and the community; and
- the need to set shorter-term developments within the longer-term framework proposed based on reporting against standards.

Such a governance structure might have the following elements:

- a sponsor – perhaps the Minister for Education and Sports – a leader who is committed and accountable and who strongly and visibly supports the enhancement of Uganda's assessment system;

- a steering committee that is accountable to the sponsor for the success of this process (the steering committee will have members drawn from across the wide range of key stakeholders)
- an implementation manager who is accountable to the steering committee;
- an implementation team that is accountable to the manager for working with the various agencies involved in Uganda's assessment system to ensure that the various elements of the road-map are implemented in a timely and effective way (the implementation team will include and have access to technical advisers); and
- a senior technical adviser who is available to the manager and attends three-monthly steering committee meetings to provide independent advice on quality issues and risk mitigation actions.

The implementation manager and team must be located somewhere. It is clear from the feedback that they should not sit within one of the agencies directly involved with assessment, exams or standards. The team could therefore sit within one of the central departments of the MoES (e.g. planning or basic education) or, and perhaps more appropriately, outside the MoES.

Funding from an external donor will be required for this governance structure, as well as for the technical capacity-building required by its operations in managing and supporting the implementation of the road-map.

The following table provides a timeline illustrating the sorts of steps required in the very early and preliminary stages of implementation of the road-map, including getting agreement amongst stakeholders about priorities in identifying and developing approaches directed towards the 'thick' problems, maintaining a focus on setting standards and reviewing against them, and planning the next stages.

Table 3: Illustrative timeline for the first seven months of implementation of a five- to ten-year programme

	Month 1	Month 2	Month 3	Month 4
Programme management	First meeting of steering group – high-level outcomes for the programme agreed; governance arrangements signed off; secretariat appointed	Secretariat develops project plans, including risk registers	Secretariat develops project plans, including risk registers	Second meeting of steering group – Programme planning documents agreed and discussed
Defining standards	Agencies (e.g. central agencies, such as UNEB) tasked to develop standards	Agencies work with stakeholders to develop standards	Agencies work with stakeholders to develop standards	Agency standards for forthcoming year presented and agreed at steering group meeting
Addressing thick problems	Initial list of 'thick problems' discussed	Secretariat works with agencies and stakeholders to generate early ideas for ways of addressing identified thick problems	Secretariat works with agencies to generate early ideas for ways of addressing identified thick problems	First thick problem to tackle is agreed (for example, the impact of commercial practice tests on teaching and learning in P4–P7)

	Month 5	Month 6	Month 7
Programme management	Secretariat monitors progress against plans	Secretariat monitors progress against plans	Third meeting of steering group – progress against plans scrutinised
Defining standards	Agencies develop reports of their current work against standards	External audit of agency reports	Steering group discusses and reviews reports and audit
Addressing thick problems	System-wide work on thick problem 1 – using external resources (from Department for International Development (DFID) programme) as necessary	System-wide work on thick problem 1 – using external resources (from DFID programme) as necessary	Remaining thick problems to address agreed and ordered. Steering group identifies actions for agencies in subsequent years and determines resources required

The implementation of the processes and steps described in this road-map are designed to have the following outcomes:

- continuous improvement approaches are valued and used by agencies and supported by Government;
- 'thick' problems in Uganda's assessment system are acknowledged, identified, understood and prioritised by stakeholders, with improvements developed and implemented iteratively; and
- broad stakeholder engagement in education improvement processes.

The outputs will include:

- agreed standards for the processes and products of agencies involved in education;
- capacity-building linked with stakeholder-agreed priorities and standards;
- independent reviews of agency performance; and
- positive and negative performance incentives for agencies at central, district and local levels.

Bibliography

Acana, Sylvia, Dan Kyagaba, Amos Opaman, Omala St Kizito, Jumanyol Kennedy, and Sserunkuma Lutalo Bbosa. 2012. "The achievement of primary school pupils in Uganda in numeracy and literacy in English." Tech. rep., UNEB National Assessment of Progress In Education.

Acane, Sylvia. 2006. "Reporting results of national assessment - Uganda experience." *32nd annual conference of the International Association for Educational Assessment .*

ACARA. n.d. "Assessment and teaching of 21st century skills."

Adana, Sylvia. 2006. "Reporting results of national assessment – Uganda experience."

Allen, J. R. 2006. "Understanding the impact of certification practices – towards a systems framework." Tech. rep., Paper given to the annual conference of the Australasian Curriculum Assessment and Certification Authorities.

Allen, Reg. 2012. *Australia - Developing the enabling context for school-based assessment in Queensland, Australia.* Systems Approach for Better Education Results (SABER): World Bank.

Allen, Reg. 2006. "Understanding the impact of certification practices – towards a systems framework." Tech. rep., Paper given to the annual conference of the Australasian Curriculum Assessment and Certification Authorities.

Altinyelken, H.K. 2010. "Curriculum change in Uganda: Teacher perspectives on the new thematic curriculum." *International Journal of Educational Development.*

Alum, Joyce M. 2002. "The effect of setting Examination questions on student performance in Biology at Uganda Advanced Certificate of Education." Tech. rep., UNEB research report.

Alum, Joyce M. 1988. "The validity of UCE in predicting performance of candidates in UACE geography examinations." Tech. rep., UNEB research report.

Andrews, Matt, Lant Pritchett, and Michael Woolcock. 2016. *Scaling PDIA Solutions through Broad Agency, and Your Role.* CID Working Paper No. 315, Center for International Development at Harvard University.

Ball, Stephen J. 2012. *Global Education inc. policy networks and edu business.* Taylor and Francis.

—. 2012. *New policy networks and the neo-liberal imaginary.* Routledge.

Bank, World. 2012. "Uganda Student Assessment : SABER Country Report 2012." *Uganda Student Assessment : SABER Country Report 2012.*

Brixi, Hana, Ellen Lust, and Michael Woolcock. 2015. *Trust, voice and incentives: Learning from local success stories in service delivery in the Middle East and North Africa.* World Bank Group.

Cambridge Education. 2014. "Lower Secondary Curriculum, Assessment and Examination Reform Programme: CURASSE Phase 2 Completion Report." Tech. rep., NCDC.

Centre for Educational Research and Innovation. 2008. "Assessment for Learning: the case for formative assessment." OECD.

Centre for Educational Research and Innovation. 2008. *Assessment for learning: the case for formative assessment.* Tech. rep., OECD.

Centre for Educational Research and Innovation. 2005. *Formative Assessment: improving learning in secondary classrooms.* Tech. rep., OECD.

Chang'ach, Mary Boit and Sr. Ann Njoki and Dr. John Koskey. 2012. "The Influence of Examinations on the Stated Curriculum Goals." *American International Journal of Contemporary Research* 2 (2).

Cheng, Kai-ming. 2014. "Does culture matter? Education reforms in East Asia." *L'éducation en Asie en 2014 : Quels enjeux mondiaux ?.* Revue internationale d'éducation de Sèvres.

Connell, R. W. 1998. "Social Change and Curriculum Futures." *Change: Transformations in education* (University of Sydney) 1 (1): 84-90.

Davie, Sandra. 2015. "7 in 10 parents send their children for tuition: ST poll." *Straits Times.* 4 July.

DES. 2012. "How can we improve our teaching? A guide to evaluating and improving the quality of teaching, learning and assesssment." Tech. rep., occasional series: Evaluating and improving the quality of education.

DES. 2012. "How we inspect. A guide to external evaluation." Tech. rep., occasional series : Evaluating and improving the quality of education.

DES. 2012. "How well are our learners doing? A guide to evaluating and improving learners' understanding, attainment and wider achievements." Tech. rep., occasional series: Evaluating and improving the quality of education.

DES. 2012. "How well do we support our learners? A guide to evaluating and improving access, equity and inclusion." Tech. rep., occasional series: Evaluating and improving the quality of education.

Echazarra, Alfonso, Daniel Salinas, Ildefonso Méndez, Vanessa Denis, and Giannina Rech. 2016. *How Teachers Teach and Students Learn: Successful Strategies for School.* Vol. I OECD Education Working Paper No. 130. OECD.

Ejere, Dr. Emmanuel Iriemi. 2010. "Absence from Work: A Study of Teacher Absenteeism in Selected Public Primary Schools in Uyo, Nigeria." *International Journal of Business and Management* 5 (9): 115 - 123.

Elks, Phil. 2016. *The impact of assessment results on education policy and practice in East Africa.* UK DFID.

Finn, Chester E. 2013. "We've made progress but there's more to do." *Education Week*, 23 April.

Centre for Educational Research and Innovation. 2008. "Assessment for Learning: The case for formative assessment." Tech. rep., OECD, CERI.

Griffin, P and McGaw, B. and Care, E (Eds),. 2012. *Assessment and teaching of 21st century skills.* Springer.

Guillermo, Luis Hakim. 2012. *CURASSE, Labour Market Consultant Input Progress Report.* NCDC.

Guloba, Madina, James Wokadala, and Lawrence Bategeka. 2010. "Does Teaching Methods And Availability Of Teaching Resources Influence Pupils' Performance: Evidence From Four Districts In Uganda." Economic Policy Research Centre.

Harlen, Wynne, and Ruth Deakin Crick. 2003. "Testing and Motivation for learning." *Assessement in Education* 10 (2): 169-207.

Transparency International. 2011. *Forms and Effects of Corruption on the Education Sector In Vietnam.* Transparency International.

Kagaba, Peter. 2009. "Variations in setting format and syllabus context in language subjects and their effects on students' Performance at UCE level." Tech. rep., UNEB research report.

Kagoro, Thomas. 2008. "Assessment considerations for special needs candidates at Primary Leaving Examination, Uganda Certificate of Education and Uganda Advanced Certificate of Education Examinations." Tech. rep., UNEB research report.

Kamwine, Enid. 2012. *A competence-based teaching and learning approach in the 'O' level curriculum in Uganda.* NCDC.

Kamwine, Enid. 2015. "A summary of the findings from a study of the effectiveness of the thematic curriculum." NCDC.

—. 2012. *something.* NCDC.

Levin, Ben, and Lauren Segedin. 2011. "International Approaches to Secondary Education." Tech. rep., Higher Education Quality Council of Ontario.

Macharia, Dr. Susan M., and Nicholas S. Kiruma. 2014. "What Ails School Inspection in Uganda? Reflections, Experiences and Emerging Issues." *International Journal of Education and Research* 2 (3): 1-12.

Martin, Gayle H., and Obert Pimhidzai. 2013. *Service Delivery Indicators: Education, Health – Kenya.* World Bank.

MoES. 2012. *Annual report of the Directorate of Education Standards.* MoES.

MoES. 2011. *Skilling Uganda BTVET STRATEGIC PLAN 2011-2020.* MoES.

Moss, Connie M. 2013. "Research on Classroom Summative Assessment", chapter in *SAGE Handbook of Research on Classoom Assessment*, edited by James H. McMillan, 325-356. SAGE Publications.

Muwumba, Mukakanya Abel. 2014. "The challenges of assessing competencies and its implications on performace in national examinations in Uganda." Tech. rep., International Association for Educational Assessment Conference paper.

Najjumba, Innocent Mulindwa, and Jeffery H. Marshall. 2013. *Improving Learning In Uganda, Volume 2.* World Bank Studies.

Nakabugo, Mary Goretti. 2011. *Adapting Formative Assessment to Developing Country Contexts: Lessons from the implementation of continuous assessment in primary education in Uganda.* TSO Information and Publishing Solutions.

Napier, Nancy .K, and Vuon Quan Hoang. 2013. *What we see, why we worry, why we hope: Vietnam going forward.* Boise, Idaho: CCI Press.

National Planning Authority. 2007. "Uganda Vision 2040."

Ndagire, Ruth. 2007. "Primary Leaving Examination result as a predictor of performance at Uganda Certificate of Education examination." Tech. rep., UNEB research report.

Obong, Florence. 2011. "A comparative analysis of the strength, validity and reliability of English language paper 1 sections A and B of Uganda Certificate of Education Examination." Tech. rep., UNEB research report.

OECD. 2011. "Education at a glance – Editorial: Fifty years of change in Education." Tech. rep., OECD.

OECD/Asian Development Bank. 2015. *Education in Indonesia: Rising to the Challenge.* OECD Publishing, Paris.

Ogwang, Samuel George. 2012. "Relationship between entry qualifications and candidates' performance in the ordinary diploma in engineering courses." Tech. rep., UNEB research report.

Ogwang, Samuel George. 2009. "Suitability of the 2006 Primary Leaving Examination and Candidates' Performance." Tech. rep., UNEB research report.

Okelowange, Primo J. 2005. "Marking Reliability in Uganda Advanced Certificate of Education Examinations with reference to History Paper 3." Tech. rep., UNEB research report.

Penny, Alan, Michael Ward, Tony Read, and Hazel Bines. 2008. "Education sector reform: The Ugandan experience." *International Journal of Educational Development* (Elsevier) 28 (3): 268-285.

Pido, Saverio. 2004. "Policies and Procedures on the conduct of coursework assessment in Uganda National Examinations Board." Tech. rep., UNEB research report.

Pido, Saverio. 1998. "Reliability and validity of hte 1997 Uganda Advanced Certificate of Education Examinations." Tech. rep., UNEB research report.

Pido, Saverio. 2006. "Reliability of the conveyor-belt marking system and the traditional marking system in the primary leaving examinatios in 2003 and 2004." Tech. rep., UNEB research report.

Pido, Saverio, and Joyce M. Alum. 1994. "An investigation into the causes of poor performance in examination in literature in English in secondary schools in Uganda." Tech. rep., UNEB research report.

Resnick, Lauren B., and Larry Berger. 2010. *An American Examination System.* Center for K–12 Assessment and Performance Management.

Stroh, David Peter. 2015. *Systems Thinking for Social Change.* Green Publishing.

Thomson, Sue, Lisa de Bortoli, and Sarah Buckly. 2013. "PISA in brief: highlihigh from the full Australian report: PISA 2012: How Australia measures up." Tech. rep., ACER.

Toohig, Elizabeth. 2014. "What Can the Child Do? A Case Study of Continuous Assessment in a Ugandan P1 Class." Master's thesis, International Education and Development, University of Sussex.

Toulemonde, Bernard. 2004. " La discrimination positive dans l'éducation: des ZEP à Sciences Po. ." *Pouvoirs* 4: 87-99.

Transparency International. 2011. *Forms and Effects of Corruption on the Education Sector In Vietnam .* Transparency International.

Trewin, Dennis. 2001. "2001 Year Book Australia." Tech. rep., Australian Bureau of Statistics.

Tucker, Marc. 2014. *Chinese Lessons: Shanghai's Rise to the Top of the PISA League Tables.* Center on International Education and Benchmarking.

UNEB. 2015. "The achievement of primary school pupils and teachers in Uganda in numeracy and literacy in English."

Uwezo (Uganda). 2013. "Are our children learning? Annual learning assessment report 2012." Tech. rep.

Wall, Patrick. 2015. "Cheating allegations rise under de Blasio, continuing a Bloomberg-era trend." www.chalkbeat.org .

Addendum A: A note on the terminology used in this report

Term	Meaning	Notes
Assessment	Any process that provides a judgement about a student's achievement	Can be formal or informal
Formative assessment	Assessment intended to help plan a student's learning	Looks forward, needs to establish details about strengths and weaknesses Sometimes aligned with 'assessment **for** learning'
Summative assessment	Assessment intended to describe what a student has learned	May be a single overall measure of achievement, may be used for selection purposes Sometimes aligned with 'assessment **of** learning'
Assessment as learning	Assessment that is an integral part of a student's learning	Sometimes contrasted with assessment of learning and assessment for learning
Assessment for learning	Assessment where the student is an active participant, becoming a self-regulating learner	
Assessment of learning	Summative assessment	
Certification	A formal, official and accepted record of a student's standard of learning in one or more areas of achievement	Issued by an official agency Accepted by other organisations as a document of record
Classroom assessment	Assessment that happens in the classroom	May be formal or informal, may be summative or formative or both
Continuous assessment	Assessment that is part of classroom/school activities	Not necessarily continuous, may have a formative or a summative function
Terminal assessment	Assessment at the end of period of learning	Usually summative in intent

Examination	Formal assessment against a set of stated requirements conducted by an independent agency	May be written, oral or practical examination under standardised conditions
Competence	Knowledge/skills transferable to new situations/contexts	
High-stakes assessment	Assessment that has important or major consequences for someone	As compared with low-stakes assessment: an assessment where the results have no major consequences tests may, for example, have major consequences for teachers (high-stakes) but no real consequences for students (low-stakes)

Addendum B: Consultations

We would like to thank all those many people who so generously gave us their time to help us deepen our understanding of Uganda's assessment system. These include the following. (Our apologies for any omissions or misspellings.)

Persons met with during the development of this road-map	Titles
John Agaba	Team Leader, Education, Prime Minister's Delivery Unit
Grace Baguma, Lydia Asiimwe, Vincent Funi, Bernadette Karuhanga, Angela Kyagaba, Christopher Muganga, Mathias Mulumba, Gertrude Namubiru and Gabriel Obbo-Katandi	Director and staff members at NCDC
Mathew Bukenya, Dan Nokrach Odongo, David Weerhe, James Turyatemba, Peter Wakabi, Dr Pido, Florence Buyinza, Dan Kyagaba	Executive Secretary (former and present) and staff members at UNEB; Director and staff members at NAPE
Godfrey Dhatemwa	Commissioner, Planning and Policy Analysis
Charles Isabirye and staff	St Lawrence Primary School, Kigoowa
Dr Rose Nassali Lukwago	Permanent Secretary, MoES
Innocent Mulindwa, Elizabeth Ninan	World Bank
Professor John C Munene	Makerere University Business School
Mr Mutandziwa, Francis Atima, Kedrace Turyargyenda	Director and staff members at DES
Dr Goretti Nakabugo	Uwezo Twaweza Lead and Manager, Uwezo
Derrick Namisi	Planning and Policy Analysis staff member
Margaret Nansasi Nsereko and staff	Commissioner and staff members at TIET
Joyce Ngaiza, Catherine Kawamara, Margaret Arica, Faith Nampeera and Maureen Katende	Staff members at DFID Uganda who have children of school going age
Dr Daniel Nkaada	Commissioner Basic Education
Dr Yusuf Nsubuga	Director Basic and Secondary Education
Onesmus Oyesigye	Executive Secretary UBTED
Dr Savario Pido	Dep. Sec. Research and Data Department, UNEB
Rosemary Rugamba-Rwanyange and Emmi Pakkala	UNICEF
Ambrose Ruyooka	Deputy Project Coordinator Uganda Teacher and School Effectiveness Project

Emmanuel Ssewankambo	Overseas Development Institute, Budget Strengthening Initiative
James Tweheyo (General Secretary), Wajega Juliet Sasagah and Bagume Filbert Bates	General Secretary UNATU and staff UNATA

Annex A Analysing UCE and PLE data

UCE examination data sets provide information about participation patterns relevant to policy and practice.

An analysis of item level data in a sample of PLE examinations shows how these examinations are focused on selection rather than on standards.

This annex has two parts. The first examines a dataset of UCE results. The second analyses four PLE examinations, item by item, using item marks entered from a sample of marked scripts.

A.1 Analysing a UCE dataset

We had a sample dataset of 200,000 randomly selected students from UCE 2015, providing basic demographics (gender, birth year, district, centre), results (in divisions from 1 to 9 – 1 is the highest) in 49 subjects, a UCE division result (1 to 4, 1 is the highest), a PLE division result (1 to 4 plus U), a division result in each of PLE English, Mathematics, Science, and Social Studies with Religious Education, together with the year that each student took the PLE.

A comparison of the distributions of UCE results in this sample dataset (see the table 1 in the appendix) with the published distributions indicates that the sample is probably representative of the full dataset – which itself represents about half of the age-cohort of about 600,000.

A comparison of the PLE divisions with the typical distributions for the full PLE candidature indicate, as might be expected, that the UCE candidates are much more likely to have better PLE results: the weaker PLE candidates tend not to go on to UCE level. We do not know if they finish at P7 or at some stage before UCE, or if they are not entered for UCE, although it appears that the number of students taking BTVET examinations is a very tiny proportion (less than 1%) of the age group.

There are many more males than females in this dataset – suggesting that females are more likely to not reach UCE level, since it appears that there are slightly more females than males in this age group. The participating males have higher PLE results (proportionately more division 1s) than the females.

Of the 49 subjects, there are seven with large enrolments (nearly the total participating cohort): English, Mathematics, Geography, History, Chemistry, Physics and Biology. There are two with enrolments slightly more than half the the total cohort: Christian Knowledge and Agricultural Practices. Of the remaining 40 subjects, most (32) have enrolments less than 2% of the total cohort and eight have enrolments between 2% and 13% of the total cohort.

All of the seven large and two medium-sized subjects show tiny proportions of candidates with the higher results (divisions 1 and 2). Eight of these nine subjects have a distribution of PLE results comparable with each other – Agricultural Practices has, on the basis of PLE results, a weaker cohort (one that does less well at PLE).

These observations are consistent with what we see from a review of examination papers and the examiners' reports: these are examinations that tend to give a much greater emphasis to items that discriminate rather than to items that provide opportunities for most students to show what they know and can do (the 'competencies' described in the syllabuses).

In the small cohort subjects there are instances – such as French – where there is a high proportion of division 1 results. In nearly all these cases, the participating cohort is, in terms of PLE

results, a high achieving cohort. The French examination paper we reviewed (there are also oral and written examinations), however, also shows a pattern of comprehension items that emphasise lower level comprehension skills – items where the correct answer can be found from one sentence in the passage.

There are, however, cases, such as Building Practices, where a high achieving PLE cohort has no division 1 results. These cases are not inconsistent with our earlier observation that the examinations emphasise discrimination more than competence, although there may be other factors involved, including the very small numbers of students.

We do not have item by item scores for the UCE examinations. Such a dataset would allow us to check these observations and provide further illumination.

Selection for secondary school is one of the key functions the PLE was designed to support. A key perspective on the validity of the PLE is therefore the extent to which PLE results are a predictor of UCE results. We have made an initial exploration of this perspective, using mixed effects modelling. In this initial work, we have treated the division results as if they were an interval scale (which they are probably not). In follow-up work, it would be better to use the underlying scores (which we do not have), or to model them as ordered categories.

We have run 10 analyses, modelling UCE division and results in the nine UCE subjects with large numbers of students. Details of these analyses appear in the appendix.

We included in each model, as so-called 'fixed effects', the student's birth year, gender, PLE year and PLE results in English, Maths, Science and SST. These are, of course, not effects in the sense of causes. They are attributes that are associated, to a greater or lesser extent, with UCE results. We also included, as so-called 'random effects', a code representing the UCE examination centre (it could be interesting to use the PLE examination centre).

The results of these preliminary analyses indicate that there is in most cases a clear association between both UCE division and results in these subjects (as outcomes) and the student's age, gender, the year in which the student took the PLE and the student's results in PLE subjects. The exception is that PLE English results are, once all these other factors have been taken into account, **not** strongly associated with UCE result in Physics, History and Biology.

Without exception, there are evident differences between some examination centres once all these other factors have been taken into account. Further exploration is needed to understand the meaning and significance of these differences, which could be caused by one or more of several factors, such as instruction, candidates' expression, language background, resources, school size and so on. There is a good discussion of the range of possible issues in MoES 2012, pp. 12–15.

Information from the sorts of analyses illustrated here is potentially very useful in understanding realities of Uganda's assessment system – the impact on students' results of a range of factors outside their control, such as background language, location and resources.

We observe from these results that, when all other factors have been taken into account, students with better UCE results are more likely to:

- be younger;
- be male (except in English, when they are more likely to be female);
- have taken the PLE more recently; and
- have better results in PLE Mathematics, Science and SST.

PLE English results are positively associated with UCE results, except in Physics (no association) and Chemistry and Mathematics (negative association – even after gender has been taken into account).

Overall, therefore, we see an association of PLE results and UCE results that is consistent with the PLE's use as a selection device or as an indicator of consistency between the sorts of examination skills needed at both PLE and UCE level. The value of this kind of predictive validity, of course, rests on the extent to which the PLE examinations themselves are valid assessments of the skills expected by the curriculum – see below and in Annex D for a more detailed discussion of the differences between the official curriculum and the practices of assessment.

However, this overall pattern masks the clear reality that some who do well at PLE do badly at UCE, and some who do badly at PLE do very well at UCE. For example, about one in three students with division 1 in PLE Mathematics have division 3 or worse in UCE Mathematics.

Additionally, we do not know how those who do not reach UCE level would have performed. We do know, however, for example, that one-in-four students with a PLE result of 8 or 9 in English achieve a division 3, 4, 5, 6 or 7 result in UCE English: that is, a low PLE result is not an infallible predictor of failure four years later.

The analysis given here illustrates the capacity of similar data-based reviews to support developments in policy and practice by providing an evidence base for refining understanding of the sorts of patterns that practitioners often intuit from their experience. Turning such reports into action is a different challenge.

A.2 Analysing PLE data at item level for 300 scripts in each of four subjects

As with the UCE, item scores at PLE level are not recorded electronically. We were able to obtain a sample of 300 scripts in each PLE subject, to record item scores for each of these scripts and to analyse these datasets. We suggest that this sort of process would add considerable value to UCE and PLE examiners' reports, and would help support future item development with more systematic information about what has and has not worked well in the past.

The sets of scripts were selected to give an indication of the *range* of student responses across the full range of performance. The distribution of total scores on these sets of scripts is therefore **not** the same as, or even similar to, that found in the total scores.

We do not have the distributions of raw scores in PLE, only the distributions of division results in PLE subjects discussed above, but these and the relationship between the PLE division results in this dataset suggest that the full dataset shows strongly skewed raw scores, with a very small proportion at the very top. This is consistent with the history of the PLE as a high-stakes selection examination.

Marks handwritten by markers on individual items on each script were entered into a spreadsheet. This gave a set of marks[16] that were analysed using item response theory (IRT) techniques. Using IRT means the dataset is analysed to identify information about items in a way that means that information about items is, unlike in classical test statistics, intentionally **not** dependent on the information about persons. Basic IRT models are based on an assumption that a single latent trait

[16] The English marks included some half marks – for example, where the instructions told students to insert the correct word from a list, students who copied the word without capitalising it were likely to get a half mark. For the purposes of the analyses given here we ignored these half marks.

(or ability) is measured by a test. According to the model, the probability that a person will get an item right increases with ability.

Even using the restricted sample of 300 sets of marks, IRT lets us look at the essential properties of items in these tests.

IRT requires that items are locally independent – the chance of getting any item correct is dependent on ability, not on getting another item correct. We have noticed that some items in some of these tests appear, from their content, to depend on others.

The key properties of items we want to know about are their *difficulty* and *discrimination*:

- one item is more difficult than another if the IRT model shows that higher ability is required to have the same probability of getting the item right; and
- an item has a high discrimination if it correlates well with the overall ability (latent trait) being assessed – high ability persons are more likely to get the item right and low ability persons are more likely to get the item wrong. A discriminating item distinguishes better amongst persons with similar (but different) levels of ability.

We also want to know if any items are *redundant*[17] – that is, are there sets of items that essentially provide the same information? We want different items to tell us about different aspects of the overall ability scale (latent trait), rather than being essentially repetitions of a single facet of this ability. A very high test reliability index usually indicates the presence of a lot of redundant items.

In the analysis of each PLE subject (English, Mathematics, Science and SST) we:

- fitted one-parameter and two-parameter IRT models, using the TAM package[18];
- obtained estimates of test reliability;
- obtained fit statistics for items (to tell us how well items fit the model);
- produced Wright Maps – charts that put items and persons on a single (logistic) scale. These figures let us visualise how items are spread across the scale (there is a fuller discussion of these maps below and in the technical appendix of tables and graphs); and
- used estimated item parameters for difficulty and discrimination to identify possibly redundant items.

Complete tables and graphs are provided in an appendix.

In each case, the estimated test reliabilities are all very high, possibly suggesting the presence of redundant items (and hence an assessment of an ability that is narrower than that defined in the curriculum).

[17] This is a technical term in psychometrics. See http://www.rasch.org/rmt/rmt143a.htm for a brief summary.
[18] Thomas Kiefer, Alexander Robitzsch and Margaret Wu (2016). TAM: Test Analysis Modules. R package version 1.17-0. https://CRAN.R-project.org/package=TAM

Table 4: Reliability for PLE tests

Year	Subject	Grade	# items	# observations	Reliability
2015	English	P7	100	316	0,982
2015	Maths	P7	44	316	0.968
2015	Science	P7	100	320	0,984
2015	SST	P7	97	320	0.979

A.2.1 PLE English

The Wright Map for this test follows. The distribution of the scores of the sample students (which, as noted, are not the same as the scores in the full dataset) is shown as a histogram on the left. The relative difficulty of each item (in this test, all items were scored 0 or 1, as discussed earlier) is shown by a small red triangle. This point (called a threshold) represents the ability at which, according to the model, a person is as likely as not (0.5 probability) to get the answer right.

It is easy to see that there are some very easy items (for example, item 21 (which requires students to put into alphabetical order four short words beginning with the same letter) and some very difficult items (including items 42 and 50 – these two items form part of items 30 to 50, which are variations on a task requiring students to rewrite sentences according to specified instructions). Item 42 appears to be particularly difficult because it requires students to use *about whom*, a construction that is disappearing in much contemporary spoken and written English. Item 50 is difficult because it requires students to use *had had* and a conditional *would have* in forming one sentence from two sentences expressing negatives in past events. Item 27 is very difficult, requiring the student to realise that the opposite of *bank deposit* is *bank withdrawal* (it seems likely that some who find the right word will lose the last syllable, writing *withdraw*). It is followed by a sequence of four items that are progressively easier, starting with seeing *aunt* as the opposite of *uncle* and then rearranging words into simple sentences.

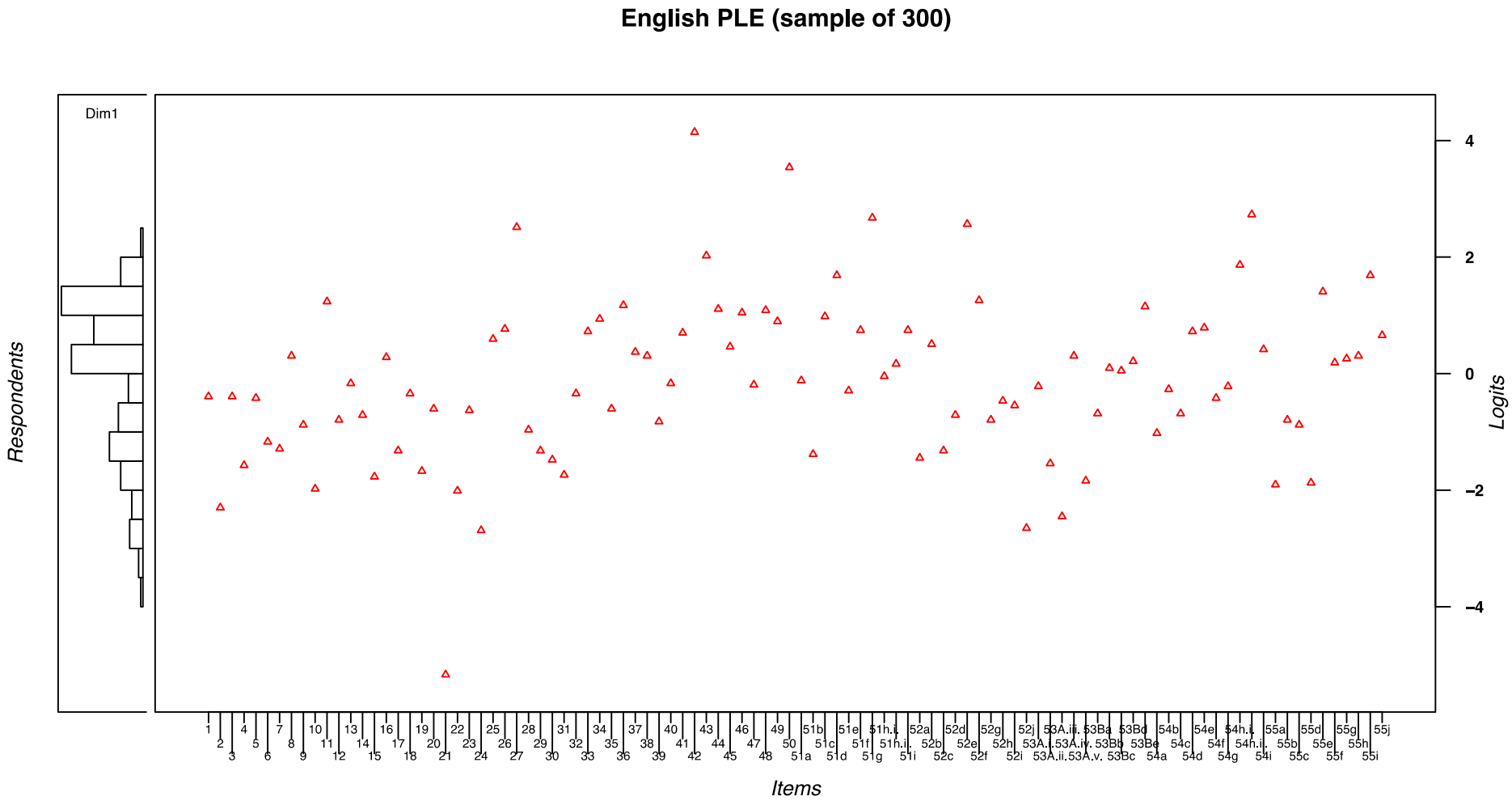
Most of the items exhibit good discrimination. Item 53a(i) (where the student has apparently only to select the word *headteacher* as the only word in the list provided that could be the person or title in an address on a letter but where a capital letter is apparently required) has low discrimination. Item 5 (fill in the blank in *Our headteacher does not want children who …. against the wall)* has high discrimination.

There are several apparently redundant items, ones with similar difficulty and discrimination. For example, items 54b and 54g both require the student to write a sentence based on finding a single fact in a passage.

Most of the features discussed above reflect the emphasis of this examination on discrimination – an examination focused on selection rather than on standards.

The following figure shows the Wright Map for PLE English.

Figure 3: Wright Map for PLE English



A.2.2 PLE Mathematics

The Wright Map for this test follows. The distribution of the scores of the sample students (which, as noted, are not the same as the scores in the full dataset) is shown as a histogram on the left. The relative difficulty of each step in each item (in this test, items were scored on a range of scales – 0, 1 on some items, 0, 1, 2 , 3, 4, 5 on three items) is shown by symbols coloured (in order) red triangle, black circle, blue cross, yellow x, pink lozenge and green inverted triangle. These points (called thresholds) represents the ability at which, according to the model, a person is as likely as not (0.5 probability) to achieve that score or higher.

The overlap or near overlap of thresholds on some items suggests that the part marks for these items do not represent real differences. Thus, for example, it appears that students get 2 marks on item 1 or no marks – few get part marks.

Items 13, 21a, 21b, 25a, 25b, 26b, 29a, 29b and 32a had maximum scores that appear to be data entry or marking errors. These have been corrected. Item 25c has five cases of scores more than 2 – these have been left in as they may reflect real scoring.

Items 23 a and 23 b (conversion of currencies) appear to be very difficult: it is possible to make any one of a wide number of errors, including choosing the wrong column (buying or selling), choosing the wrong row (the wrong currency) and making arithmetical errors.

There are, however, several items that are not independent of each other. For example, an error in the first part of item 21 or item 26 is very likely to lead to an error in the second part: almost everyone who gets 21a correct gets full marks on 21b.

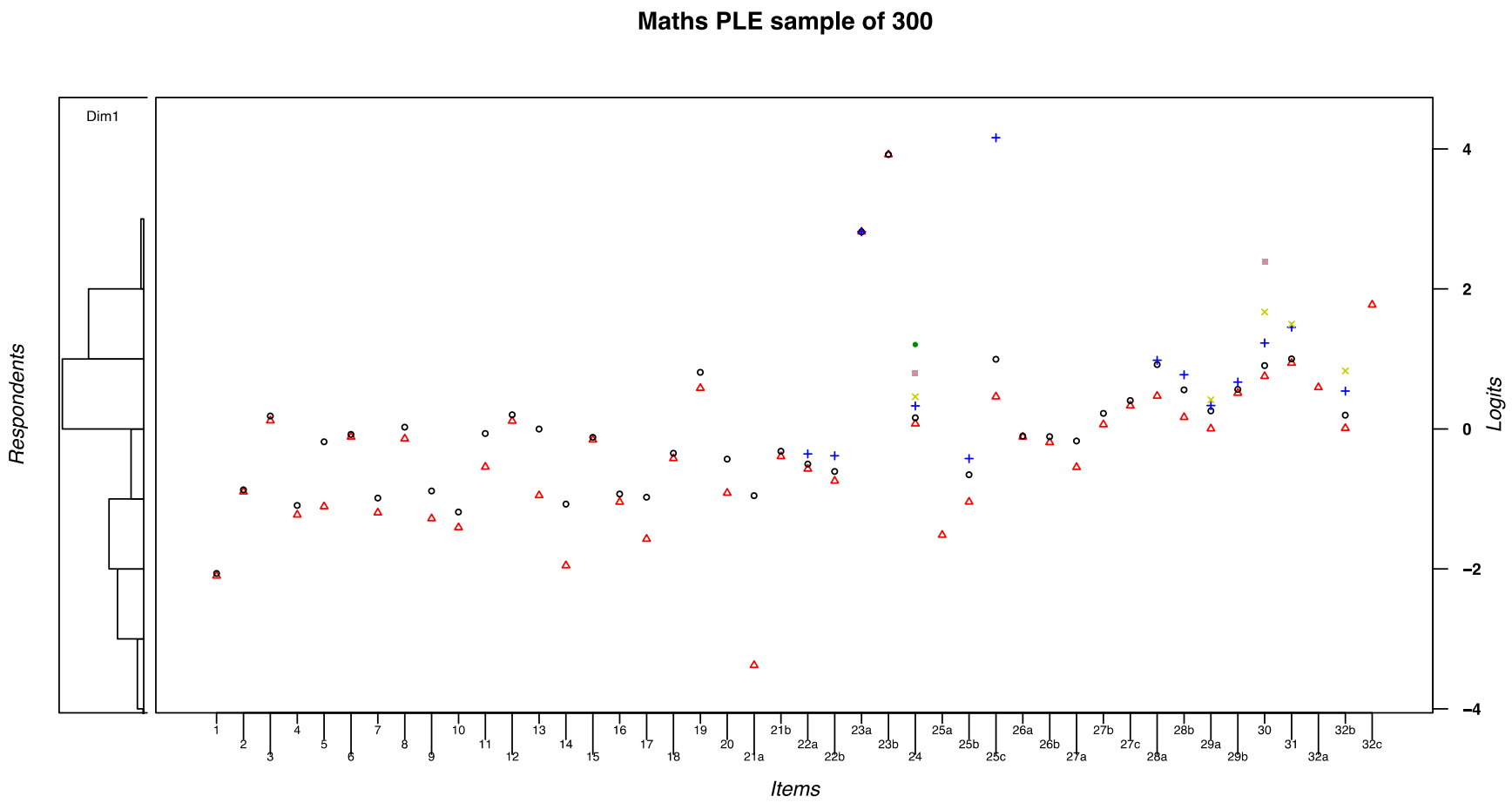
Most of the items exhibit good discrimination.

Most items appear to be of moderate difficulty, so that students with limited knowledge are less likely to show what they know and can do (their 'competencies'). High achieving students will get high scores for consistent accuracy and following the required rehearsed steps, rather for solving conceptually more difficult problems.

Most of the features discussed above reflect the emphasis of this examination on discrimination – an examination focused on selection rather than on standards.

The following figure shows the Wright Map for PLE Maths.

Figure 4: Wright Map for PLE Maths



A.2.3 PLE Science

The Wright Map for this test follows. The distribution of the scores of the sample students (which, as noted, are not the same as the scores in the full dataset) is shown as a histogram on the left. The relative difficulty of each item is shown by red triangles – in this test all items are scored 0 or 1. These points (called thresholds) represents the ability at which, according to the model, a person is as likely as not (0.5 probability) to get that item correct.

Item 1, in comparison with items 2 and 3, is apparently a very easy item – although all are basic recall questions (can you remember what you have been told in class?). Part of the reason for this may relate to language issues – the correct answer to item 1 is 'natural immunity' (which a student who initially wrote 'polio' and crossed it out was able to spell) whereas the correct answer to item 2 is 'pulmonary artery' (two unfamiliar and difficult to spell words – one student's incorrect answer was 'purunmanare vessel', showing an imperfect recall). The correct answer to item 3's instruction to 'name the energy resource obtained from animal dung and urine' illustrates the idea that these items test recall: the 'correct' answer is 'biogas' not 'natural energy'. Like these, many items are of the form 'state' or 'name'.

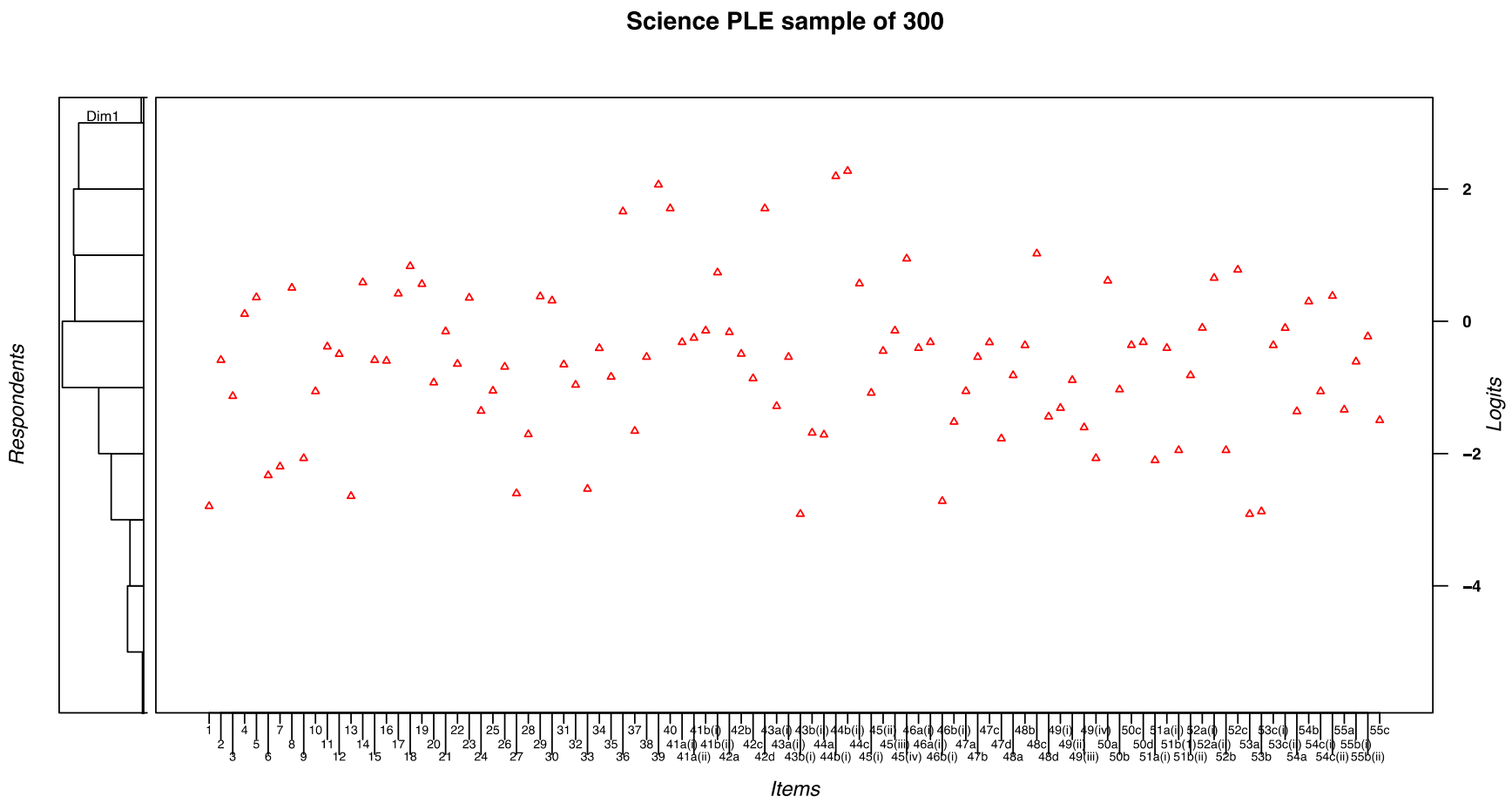
Items 12 (easy), 14 (moderately difficult) and 44b(i) and 44b(ii) (very difficult) illustrate the ways in which the items tend to test recall rather than understanding. Item 12 asks 'why is water obtained by the distillation method **not** good for drinking?'. The required answer is that it lacks mineral salts, although in comparison with ground water (or sea water) distilled water is in fact 'good for drinking', especially if trace minerals are found in food sources. Item 14 asks 'which property of magnets enables a magnetic compass to work?'. Students who got this item correct wrote (with slight variations) 'a freely suspended magnet rests facing in north south direction', suggesting recall rather than understanding of the relationship with the earth's magnetic field. Item 44b asks 'what happens to the rib cage during the processes of breathing in and out'. Correct responses refer to the rib cage going both upwards/downward and outwards/inwards – stating only one of these features is incorrect.

Most of the items show good discrimination (indices above 1). Item 40 does not. This item says 'state one reason farmers stake tomato plants'. We saw examples where 'it does not grow well' (written by a low scoring student) received a mark but 'to get support' or 'to give them extra support' (written by higher scoring students) did not.

Most of the features discussed above reflect the emphasis of this examination on discrimination through testing recall – an examination focused on selection rather than on standards or understanding.

The following figure shows the Wright Map for PLE Science.

Figure 5: Wright Map for PLE Science



A.2.4 PLE SST

The Wright Map for this test follows. The distribution of the scores of the sample students (which, as noted, are not the same as the scores in the full dataset) is shown as a histogram on the left. The relative difficulty of each item is shown by red triangles – in this test all items expect for item 49 are scored 0 or 1. These points (called thresholds) represents the ability at which, according to the model, a person is as likely as not (0.5 probability) to get that item correct.

Item 1 is, according to the data, very easy and it seems plausible that the only time a question of the form 'name any one domestic animal kept for milk' would present any challenge in a society that uses milk is to a student without a command of basic English.

For item 49, which requires students to fill in blanks in a table relating countries, political leaders and political parties, it appears easy to get one right (possibly matching Nelson Mandela with South Africa), and of only moderate difficulty to get all four correct.

Most items are of relatively similar difficulty. Item 47c is the most difficult. This item asks the student to name a climate shown in a rainfall and temperature distribution. It may be safe to assume that students were expected to identify this from recall of a classroom presentation as a winter rain/summer drought climate type, perhaps naming it as 'Mediterranean'. This item, unsurprisingly, has a high discrimination index. Item 47d is, apparently, an easy item. It asks the student to 'mention any one economic activity suitable in this area'. Apparently, the expected answer is some variation on themes of pastoralism. That is, although it is not specified, responses are required to take into account only the information about rainfall and temperature, so that mining is not a possible response. This implies that students are expected to recall what they have been taught about the relationship between climate and agriculture.

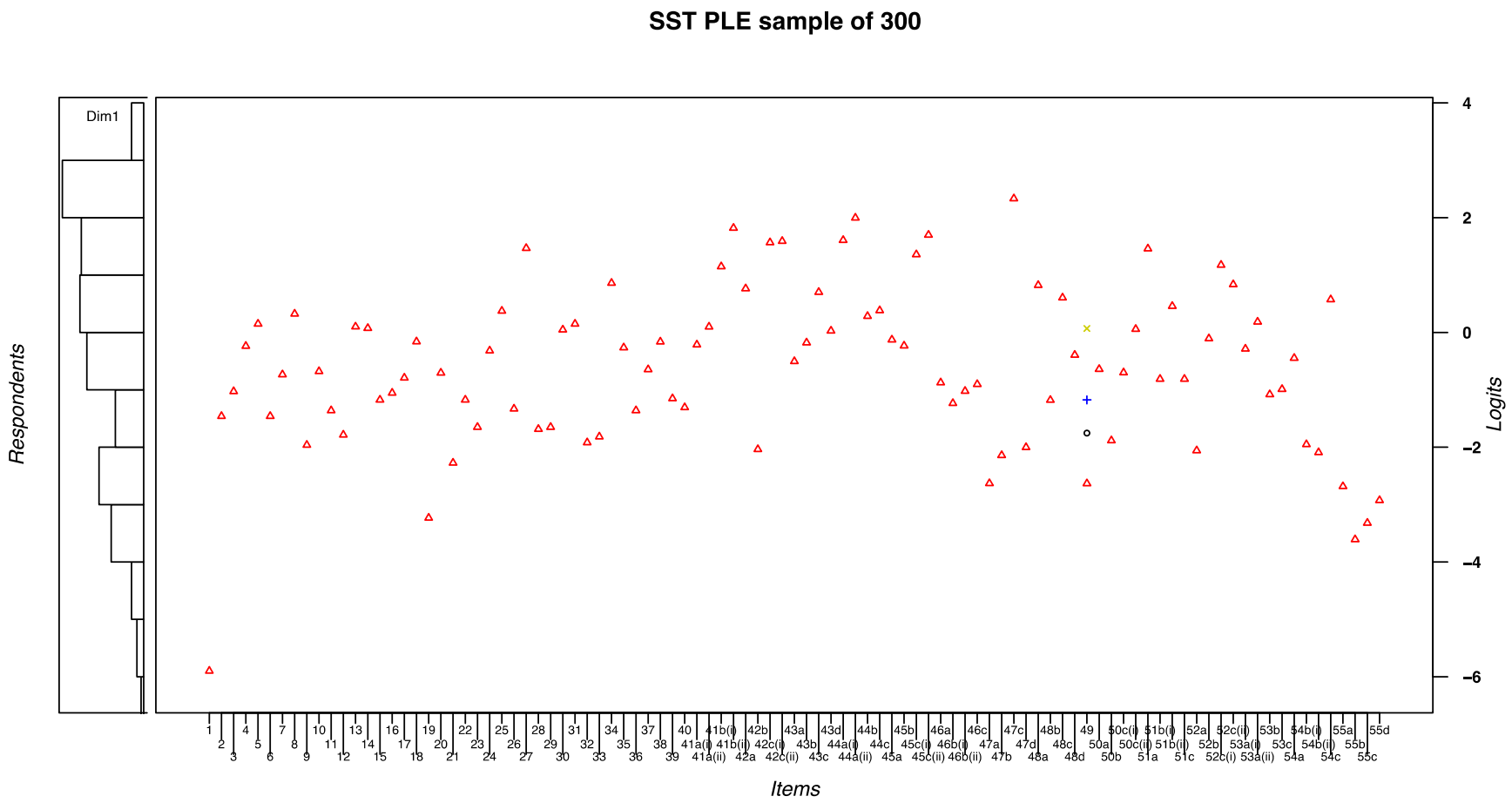
The items in this test generally have good discrimination indexes. One exception is item 27. This asks students to describe the difference between the work of a veterinary and agricultural officer. A review of a few cases suggests that this may reflect variable application of a scoring guide. We saw examples where 'a veterinary officer treats sick animals while an agricultural office gives farmers advice on how to plant' score 1 and 'a veterinary officer treats sick animals while an agricultural office advises farmers on the farm practices' scored 0.

Many of the items on this test are of the form 'name' or 'state', requiring students to recall a classroom fact such as 'name the rapids along the river Niger where Mungo Park drowned'.

Most of the features discussed above reflect the emphasis of this examination on discrimination through testing recall – an examination focused on selection rather than on standards or understanding.

The following figure shows the Wright Map for PLE SST.

Figure 6: Wright Map for PLE SST



Annex B NAPE reports and datasets

As one of the oldest national assessment of learning outcomes in Africa (1996), NAPE today draws on a wealth of experience and expertise in its development, construction and administration.

Technical reports are not of themselves an effective means to enhance Uganda's assessment system. Our analysis of NAPE identifies procedural, documentation and reporting issues similar to those discussed in a review completed over ten years ago.

This annex discusses the validity and reliability of NAPE, based on an analysis of the national reports, papers and data.

NAPE is a large-scale national assessment undertaken every year in P3, P6 and S2. A larger assessment with contextual questionnaire is administrated every three years. Uganda has carried out national assessments of educational achievement since 1996. Initially, pairs of subjects (literacy and numeracy; science and social studies) were assessed on a three-yearly basis. From 2003, the focus has been on literacy and numeracy, which are assessed annually.

We will first present the objectives of NAPE, the indicators used and the ways of reporting the data in the reports, before reviewing test content and then calculating the psychometric values of the test. This will inform the discussion of the validity and reliability of NAPE.

The following comments should be understood in the context of the limitations of this study. In the time available, we were able to review copies of the 2013 and 2015 primary papers and the 2013 S2 papers. We had data for 2012, 2014 and 2015 primary tests and 2012, 2013 and 2014 S2 tests.

We appreciate the feedback we have had that has helped us to improve this draft. Any errors, misrepresentations and infelicities remaining are our responsibility,

B.1 Review of NAPE 2015 report

The purpose of NAPE, as described in Acana (2006), is to: 'generate accurate information on what pupils know and can do in various curricular areas; evaluate the effectiveness of reforms in the education system; provide guidelines for the improvement of instruction and learning; evaluate the effectiveness of inputs and processes employed in the educational activities; provide guidelines on variables that affect achievement; and provide data that can be used in planning and research.'

According to the NAPE 2015 report, the objectives are to:

1. determine and monitor the level of achievement of pupils over time;
2. generate information on what pupils know and can do in different curricular areas;
3. evaluate the effectiveness of reforms in the education system;
4. provide information on variables that affect learning achievement;
5. suggest measures for the improvement of teaching and learning in schools; and
6. provide data for planning and research.

The instruments are intended to be criterion-referenced tests of literacy and numeracy, based on the national curriculum. Looked at from certain perspectives, there have been significant and major changes in the official national curriculum over the time that NAPE has been in operation. The limited changes in the format and content of NAPE over the same period may perhaps indicate a focus on relatively unchanging aspects of the national curriculum.

No psychometric information on reliability and item difficulty is provided in the published reports. Recommendations are general and focused on teaching methods. Data are disaggregated by district and by competence/skill/learning outcome. The indicators used are the percentages of pupils achieving proficiency or reaching minimum competencies.

A table of specifications informs test content by domain/strand, but not by cognitive abilities.

All items are constructed response questions – there are no forced choice items.

The levels are defined as follows:

Table 5: NAPE scale

Advanced level:	**Indicates superior performance. A pupil with this rating would have demonstrated complete mastery of the subject matter.**
Adequate level:	Demonstrates competence in the subject matter. This is the desired minimum performance level that was required of all the pupils.
Basic level:	Demonstrates competence in elementary concepts and skills. The pupil is performing at a level below his/her class.
Inadequate level:	Demonstrates competence in only rudimentary concepts and skills, and the pupil is performing far below the expected level of his/her class.

Source: 2015 NAPE report

A pupil is rated proficient if he/she reached the 'Advanced' or 'Adequate' level of proficiency.

We do not have a clear understanding of how cut-scores are determined so that these levels can be applied consistently to different tests. The process appears to draw on understanding the items and determining the total score on a test (before it is used) that experience indicates will correspond to the required standards.

The proportion of pupils achieving proficiency is fairly high: over 60% in P3 (over 95% in private schools) and over 50% in P6 (over 70% in private schools). Pupils from private schools generally do well.

Data are not reported by cognitive ability.

Data are reported at district level, with ranking and by year (trends). No confidence intervals are provided in the published reports. In similar assessments (international, such as PISA, or national, such as Ghana) with equivalent sample design, data are not reported at district level because of the small number of schools, but rather at region level. It is common practice to consider samples of 40 to 50 schools per district to calculate reasonable estimates.

Confidence intervals using appropriate techniques are an important aspect of many reports of standardised testing, providing some guidance on the interpretation of results. The published NAPE reports do not include these. Confidence intervals are not easily understood by non-technical readers but the issues they represent are. For example, a district would find it useful to know if its results are so different from those in another district that it should take some action, or if

the difference is too small to have any practical implication. There are several ways in which such information can be presented graphically rather than in complex tables. For example, some places use a 'traffic light' convention, where red signals an important difference, yellow a difference that may repay attention and green a small, probably unimportant difference.

Again, policy-makers need to know if they should make decisions based on observed year to year changes in NAPE scores, whether across Uganda as a whole or between districts. The calculation of confidence intervals should underpin the appropriate presentation of results to support this sort of decision-making.

Data could be reported at domain, or to a lesser extent at strand level, in the NAPE report since there are at least five items by domain.

B.2 Review of test content

B.2.1 Test length and duration

National large-scale assessments in other countries are often noticeably shorter than the NAPE tests. The specifications for NAPE tests and the expectation that they will cover all the relevant elements and contributory component skills listed in the curriculum leads to tests that last 1.5 hours in P3 and two hours in P6. Large-scale tests with similar overall objectives are often around 45 minutes at P3 and one hour at P6. The number of items in the NAPE tests is consistent with their duration, at about two minutes per item[19].

In 2015, in P6 numeracy the proportion of missing responses increased at the end of the test and reached 6%. At P6 literacy, the proportion of missing responses reached 8% from the middle of the test and 20% for the last fifth of the test. This indicates potential fatigue effects for this particular year. Further exploration of missing responses and fatigue issues would be helpful.

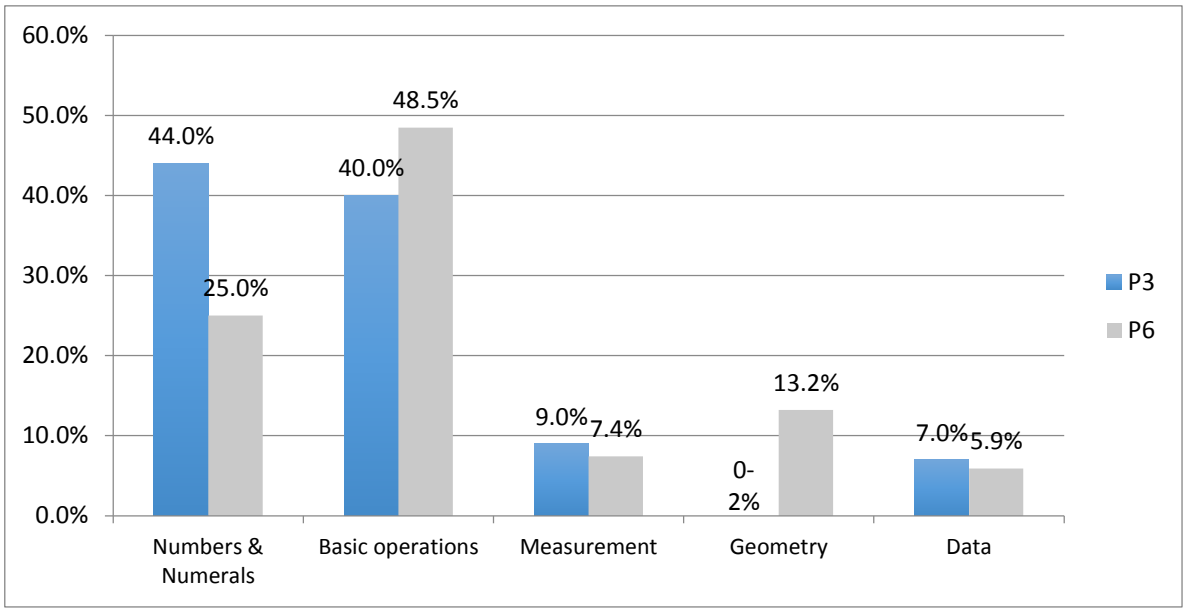
B.2.2 Domains and cognitive levels tested

B.2.2.1 Numeracy tests

The domains covered in the numeracy tests are numbers and numerals, basic operations, measurement, geometry, and data and chance. In literacy in P3 (test content specifications not provided for P6), the domains include vocabulary, writing and reading and some domains such as drawing or copying texts that are included because they are requirements earlier in the curriculum and are considered important contributory skills leading towards literacy skills. Such inclusions have an impact on the length of the tests and the extent to which they assess a well-defined construct. The texts/stories of the reading in the P3 test are very short (a maximum of six simple sentences), shorter than the ones found in EGRA (targeting Grade 2).

[1919] What constitutes an 'item' varies from one NAPE test to another. For example, in the 2012 data file for P3 literacy, item 21 appears as four sub-items, 21a, 21b, 21c, 21d, each scored 0/1, whereas in the 2015 data file it appears as one item with four score points.

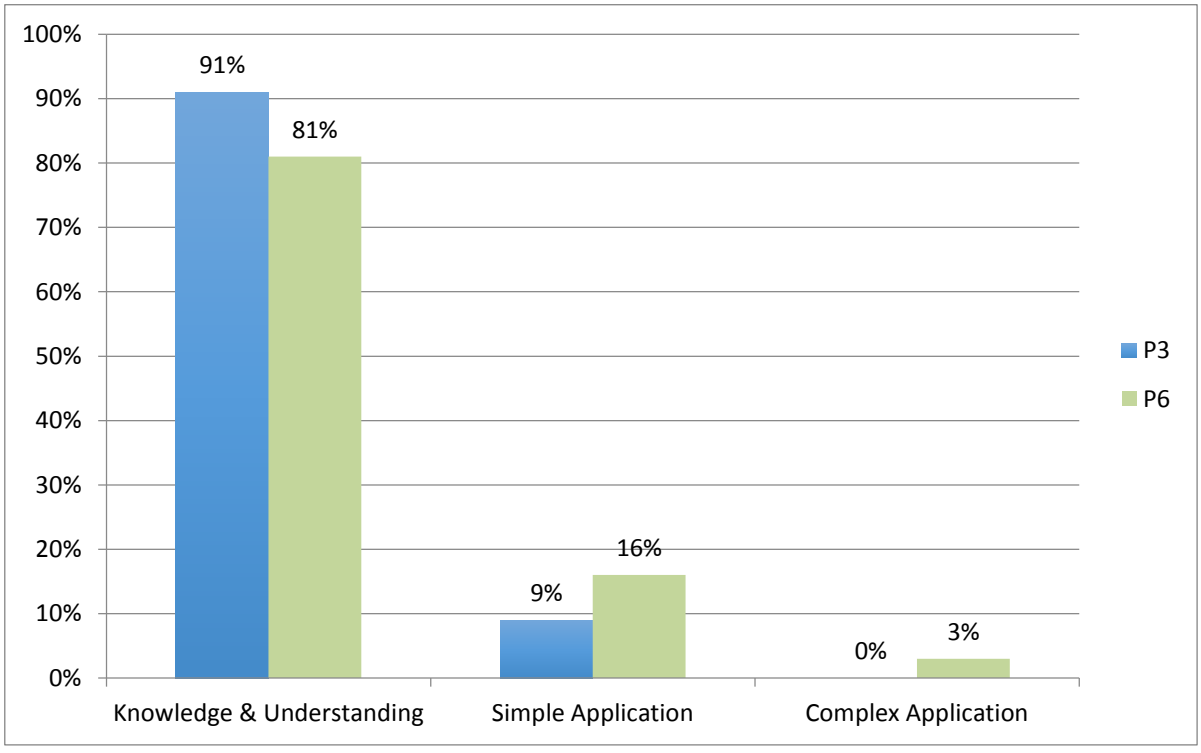
Figure 7: Share of domains by grade, NAPE P3 and P6 numeracy



Source: Author's classification

The numbers and numerals domain, which usually involves knowledge-level cognitive skills, is over-represented (44% in P3 and 25% in P6), to the detriment of measurement and geometry. Instead of none, as shown above, at P3 level two items in the 2015 test might be considered geometry – sorting and drawing shapes. The test structure by domain is not the same as in PLE mathematics, where the measurement domain is more tested. Both the P3 and P6 tests give less representation to the geometry domain than might be considered appropriate for an assessment of numeracy.

Figure 8: Share of cognitive levels by grade, NAPE P3 and P6 numeracy



The reviewer considered that the P3 and P6 numeracy tests included very few complex problems involving two steps of operation (no items in P3 and two in P6). (This depends in part on how one defines the number of steps required by an item). Application accounts for less than 20% of the tests for both grades. In comparable tests in other countries, application and reasoning (complex application) represents 50% of the test at P3 and 57.5% at P6 level[20]. These cognitive levels typically form at least 50% of the test in the SACMEQ (Grade 6) or IEA studies (Grade 4).

We understand that the construction of NAPE tests is informed by the relative amount of time allotted to various competencies in the curriculum. It is also informed by the need to gather information about what students can actually do – there is little information to be gained from a test with too many items on which many students score zero.

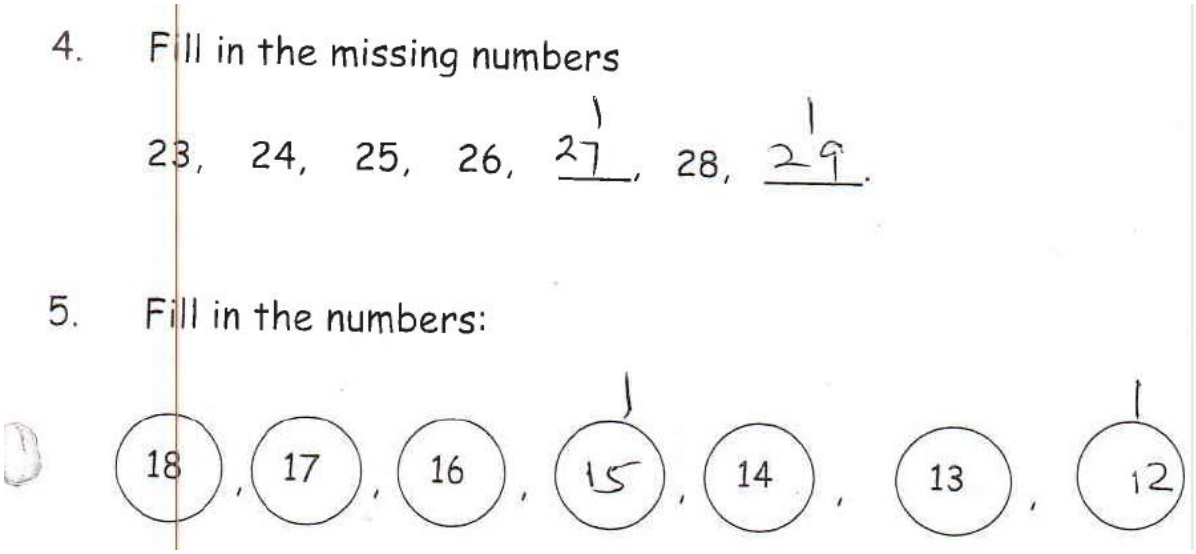
B.2.2.2 Literacy tests

A review of the items suggests to the reviewer a lack of clarity in the construct measured by the P3 and P6 literacy tests, mixing several domains or sub-skills, including skills taught as intermediate steps (such as copying) rather than literacy skills. This appears to arise from seeking alignment with the curriculum, both the curriculum at the year level tested and at the preceding levels.

B.2.3 Dependency and redundancy[21] of items

The review by item indicates the possibility of pairs of dependent items (the response to one item is linked to the response to another item) and several items that are apparently redundant[22] – i.e. that measure the same learning outcome at the same cognitive level. The presence of redundant items tends to lead to overestimates of the reliability of the test and is not the best use of testing time.

Figure 9: Illustration of two possibly redundant items at P3-NAPE 2013



While these items may be thought to be different (one is ascending numbers and the other is descending), test data (not available for the 2013 test) can be used to see if the evidence supports this. For example, there are three similar items in the 2015 test – ascending two- and three-digit

[20] Ministry of Education, Ghana Education Service, National Education Assessment Unit (2014), Ghana 2013 National Assessment Technical Report.
[21] This is a technical term in psychometrics. See http://www.rasch.org/rmt/rmt143a.htm for a brief summary.
[22] List of apparently redundant items at in 2013 P6 : 16–17; 18–21; 25–30; 47–50; 44–49 and at P3: 4–5; 4–7; 3–9; 37–38; 41–42; 44–45; 47–57.

sequences. There is a strong association of responses to these items, conditional on overall test score – that is, the data support the idea that there is some element of redundancy across these items.

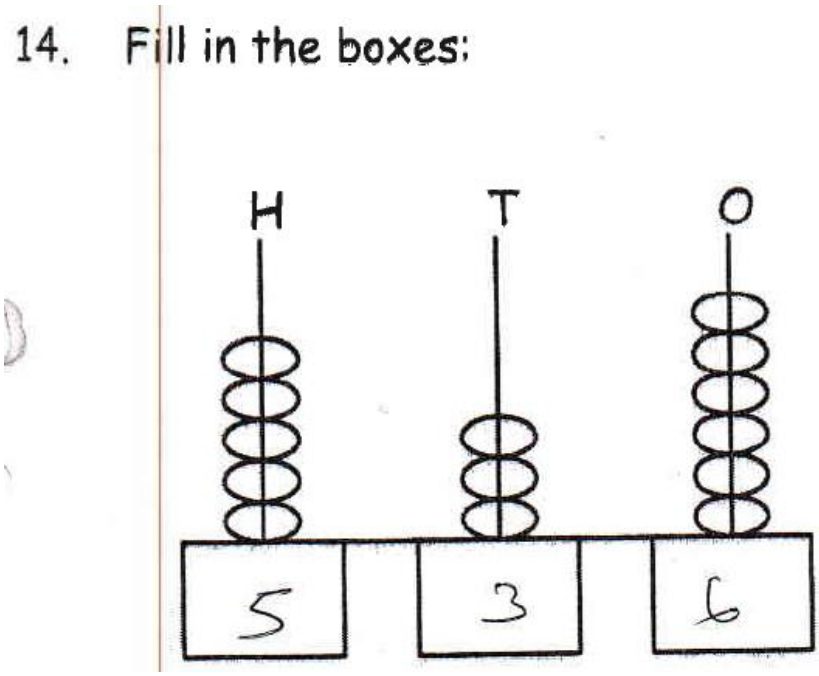
Items in the basic arithmetical operations domain differ from other large-scale tests in terms of number of digits.

These features, including the presence of possibly redundant items, may in part be attributable to a policy decision that the NAPE tests include elements from the curriculum for earlier years and items that students can be expected to get correct

In other large-scale tests it is common practice to see two- or three-digit numbers at P3 level and four- or even five-digit numbers at P6 level.

Figure 10: Illustration of easy items in the NAPE test 2013 at P3 level (comparable items in the 2015 test have facilities at about 0.7)

P3



A review of the content of the secondary tests (S2) does not show construct under-representation, or item redundancy or dependency. The English test includes language tools, reading and writing. Texts to be read are of adequate length and difficulty for this grade but are all narrative and not informational texts (of the kind that could be expected at this level).

In all NAPE tests, the stem and instructions for the pupils are not always clear or specific. For example, there may be wording that corresponds to practices in class, such as 'Work out'. Particular instructions at P3 level are, where necessary, translated.

The data analysis at item level will enlighten us further and confirm or inform the findings of the test content review.

B.3 NAPE data analysis

B.3.1 Reliability

For primary, the reliability coefficients are all very high and well above 0.90, the usual threshold for this type of tests. These high values point to potential item redundancy, as identified in the content analysis. For secondary, the reliability values are lower (0.8 to 0.93) and more aligned to expected values for these kinds of tests, since there is no item redundancy found in reviews of the test content.

Table 6: Reliability of the NAPE tests

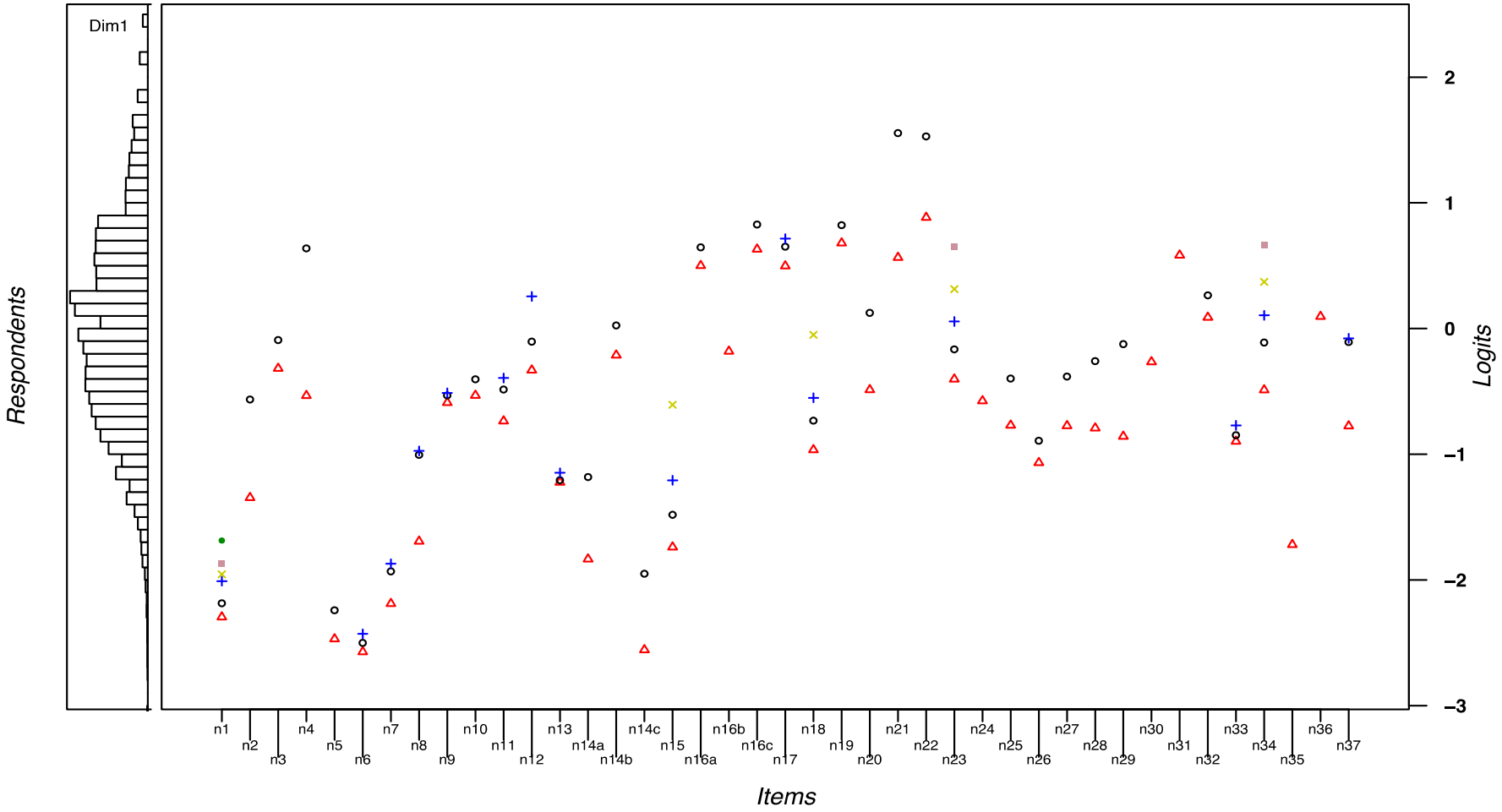
Year	Subject	Grade	# items	# observations	Reliability
2012	Literacy	P3	27	24,359	0.94
2014	Literacy	P3	56	12,155	0.97
2015	Literacy	P3	21	24,043	0.94
2012	Numeracy	P3	59	24,359	0.94
2014	Numeracy	P3	57	12,155	0.95
2015	Numeracy	P3	21	24,043	0.93
2012	Literacy	P6	27	24,359	0,94
2014	Literacy	P6	80	11,990	0.95
2015	Literacy	P6	64	22,732	0.95
2014	Numeracy	P6	66	11,990	0.94
2015	Numeracy	P6	66	22,732	0.94
2012	Biology	S2	38	19,776	0.80
2013	Biology	S2	38	21,647	0.82
2014	Biology	S2	38	19,529	0.84
2012	English	S2	80	19,776	0.92
2013	English	S2	80	21,647	0.91
2014	English	S2	80	19,529	0.93
2012	Maths	S2	47	19,776	0.87
2013	Maths	S2	47	21,647	0.87
2014	Maths	S2	47	19,529	0.85

When looking at item-level data, item-test correlations (rpbis) are all above 0.25 in most of the cases, with some exceptions.

B.3.2 Difficulty of the tests

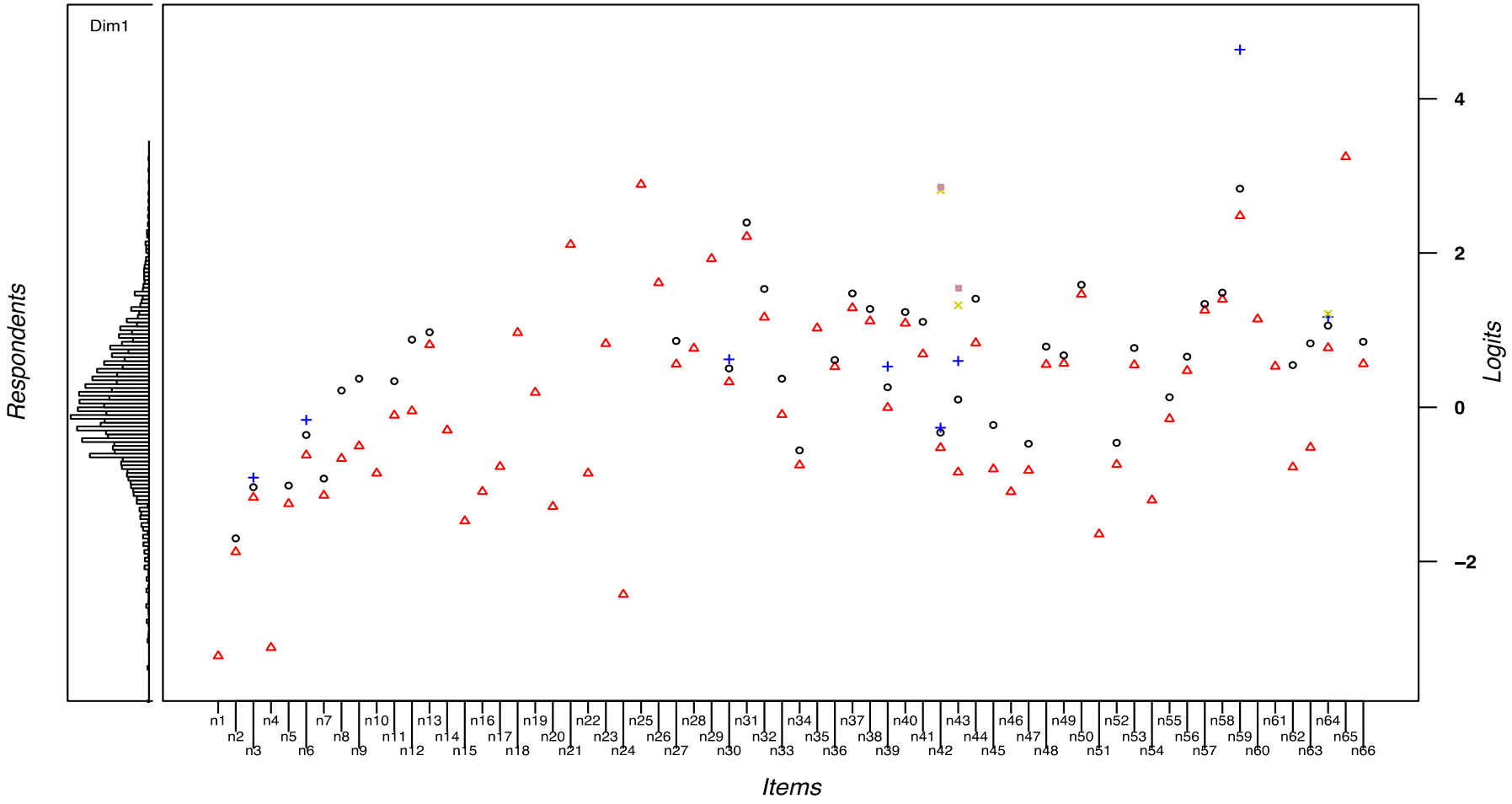
The Wright Map or Wright Plot represents pupils' abilities and item difficulty (as thresholds – see the appendix for further details, tables and graphs) on the same scale. We choose to present the most recent year (2015 for primary and 2014 for secondary). There were a number of apparent errors in this data file – scores that were outside the ranges shown in the scoring guide. These anomalies have been removed.

Figure 11: Wright Map for P3 numeracy 2015



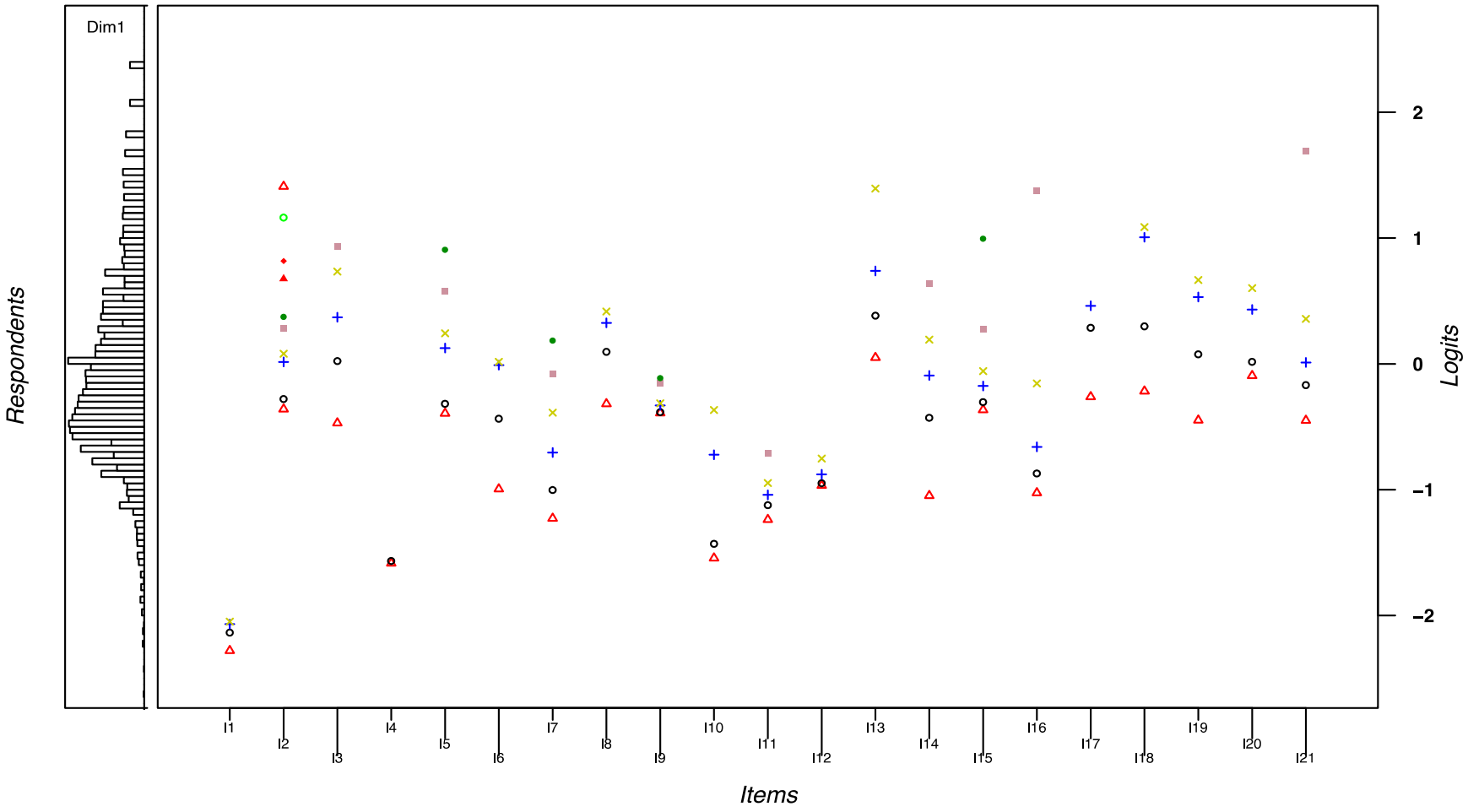
This graph shows that the test does not deliver much information for the pupils with higher abilities. Two items' thresholds are above one logit, while several items' thresholds are under -1 logit. Several items are located around -2 logits. These results indicate the P3 numeracy test is too easy and not properly adjusted to the pupils' abilities. For some items, such as items 1 and 33, the thresholds are close or even overlap, indicating that there are no real differences between these scores. Item 13 (which is similar to one of those described earlier as easy) is not only very easy but also has poor discrimination. Items 21 and 22 (basic multiplications) are the hardest items but these are only of moderate difficulty and therefore do not provide much information about higher achieving students.

Figure 12: Wright Map for P6 numeracy 2015



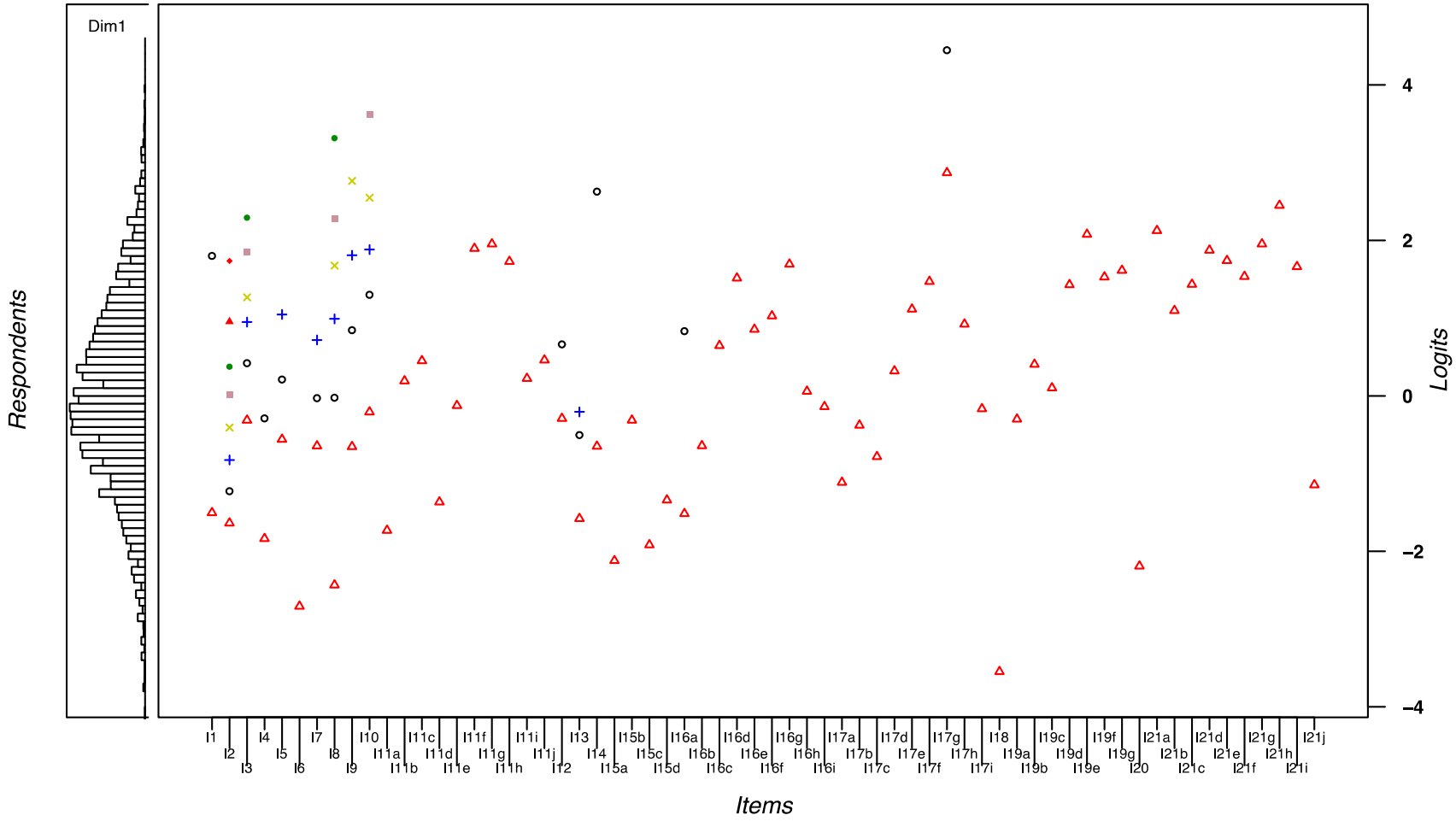
The P6 map for numeracy shows that items are almost all distributed between -2 and 2 on the logit scale, while a few thresholds correspond to higher abilities. Item 13 (representing priority rules using brackets – the actual instruction is 'use brackets to work out 6 + 4 * 5') does not fit at all well, suggesting that some students may have misinterpreted the item. Item 59 is the most difficult in the test, since very few students (11) get the maximum three marks on a multi-step item involving fractions in a word problem. There is item redundancy, since many thresholds are concentrated in the same range of ability, especially at the beginning of the test.

Figure 13: Wright Map for P3 literacy 2015



In this test the spread of item thresholds across the ability continuum for the P3 literacy test is good – noticeably better than for the numeracy tests, although there are several items where there is little difference between partial and full marks. The first item (matching pairs of pictures) clearly adds almost no information (and is one of the items that is not clearly associated with a literacy construct but rather with some skills that are considered preparatory to literacy). The second item requires naming pictures with words beginning with a given letter. A score of one (which can be seen as being of moderate difficulty) reflects one word incorrectly spelled. A score of 10 (five correct words correctly) is characteristic of very high achievers.

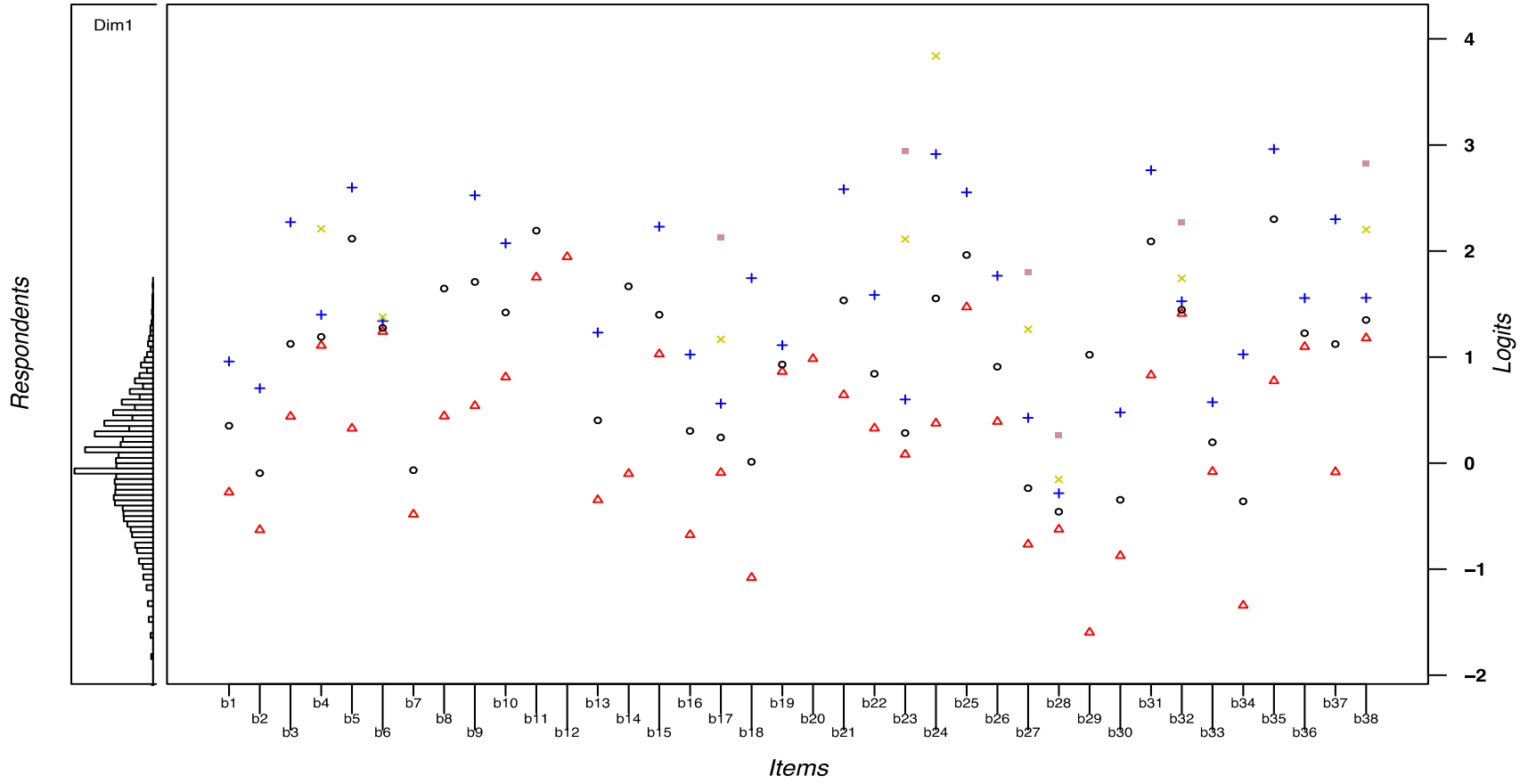
Figure 14: Wright Map for P6 literacy 2015



There is a good spread of thresholds on this test. The most difficult item is one of the nine entries for analytic marking of a letter the students are asked to write: a score of two for including reference to what the mark scheme describes as 'activities' was obtained by a few very able students. The requirements of the task did not specify what was to be included, other than a simple request. Item 10, for which full marks requires command of comparatives, is also difficult.

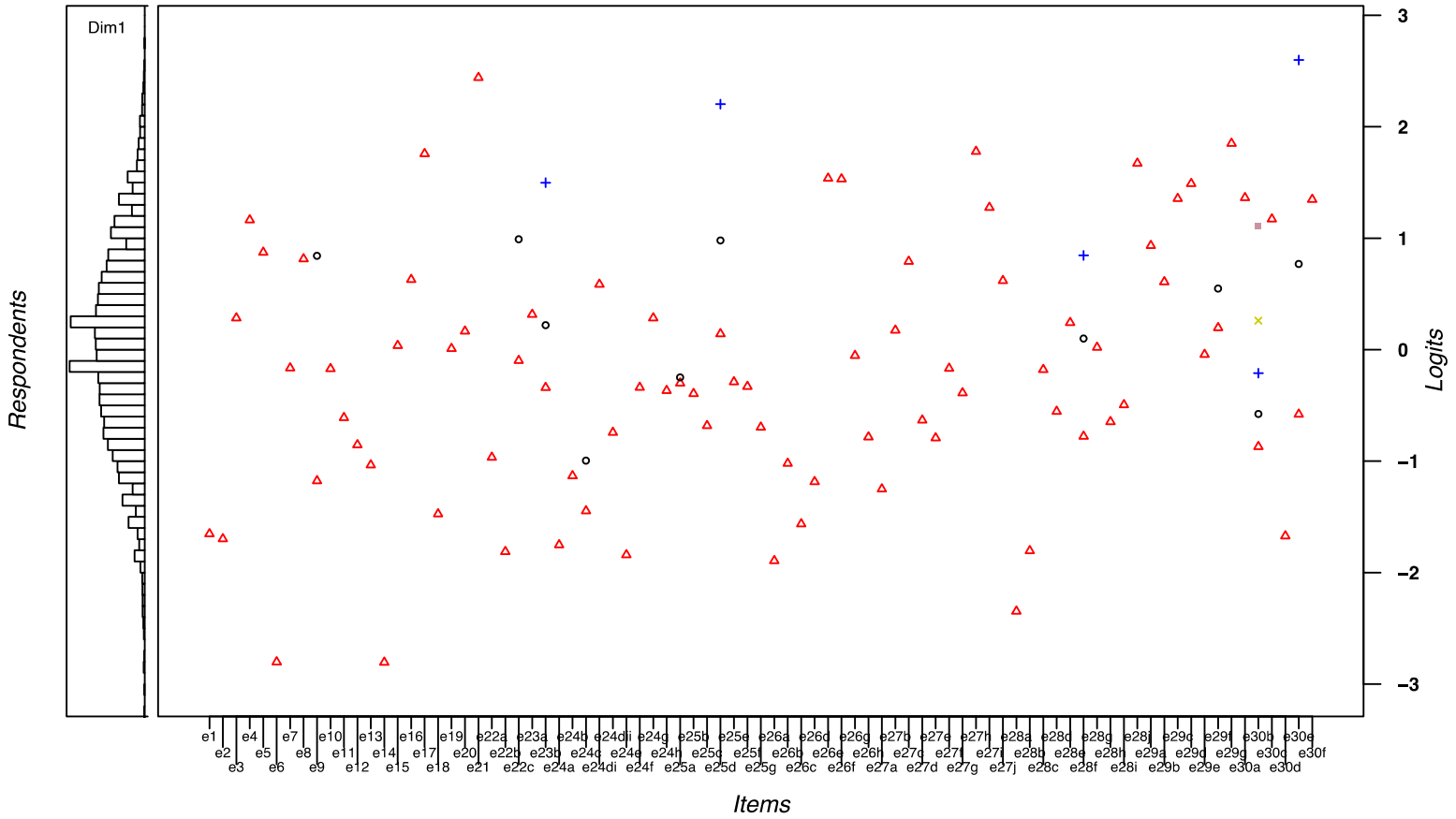
The following part presents the item-level data analysis for secondary – as noted earlier, the lower but more than satisfactory reliabilities for these tests indicates the likely absence of redundancies. Complete tables and graphs are provided in the appendix.

Figure 15: Wright map for S2 biology 2014



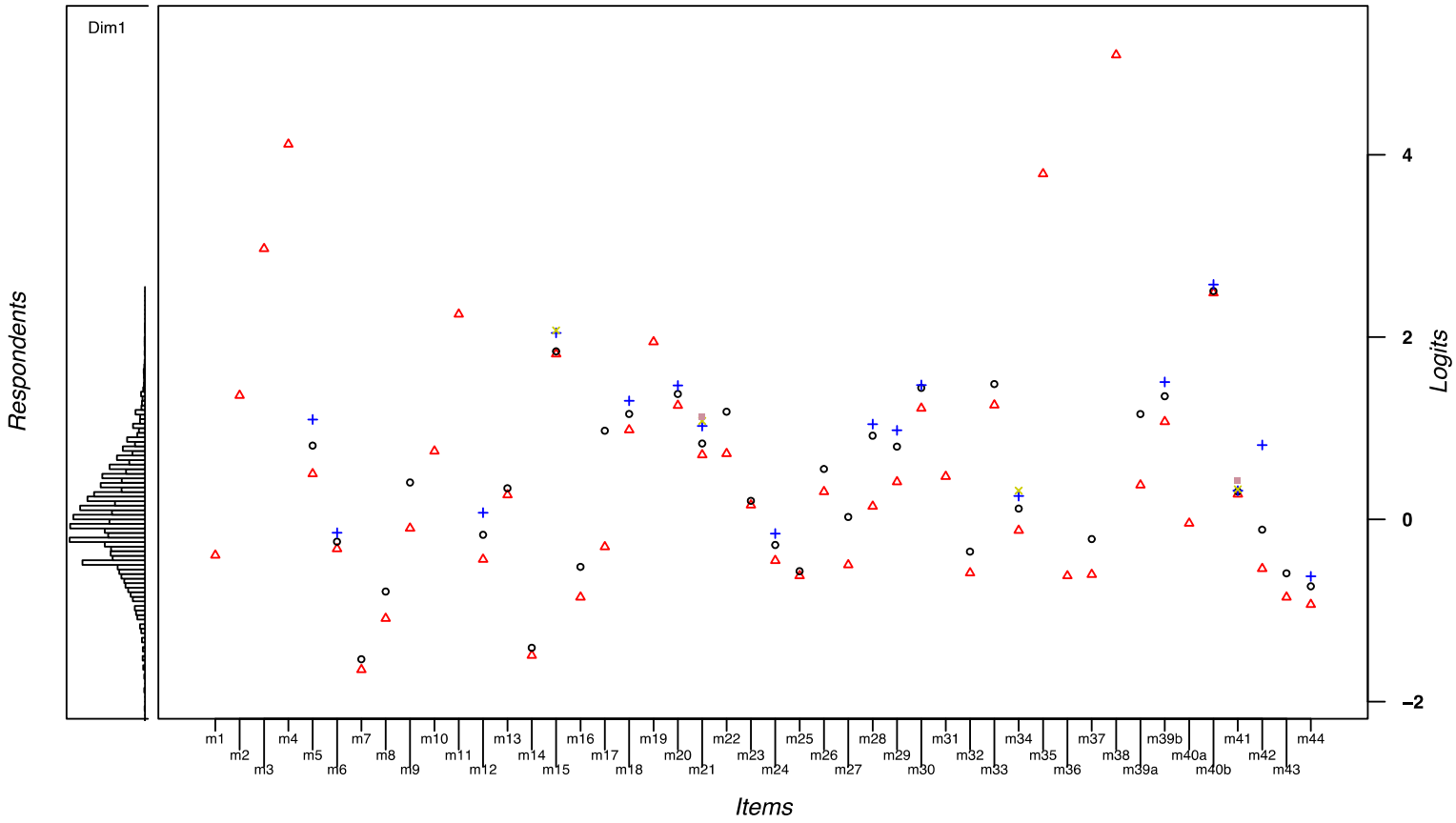
In biology, for Secondary 2, the items are well spread on the difficulty/ability continuum. There are very few misfit items or items that are located at the extreme of the continuum. This test has good psychometric properties.

Figure 16: Wright Map for S2 English 2014



In English, for Secondary 2, the items are very well spread along the continuum.

Figure 17: Wright Map for S2 Maths 2014

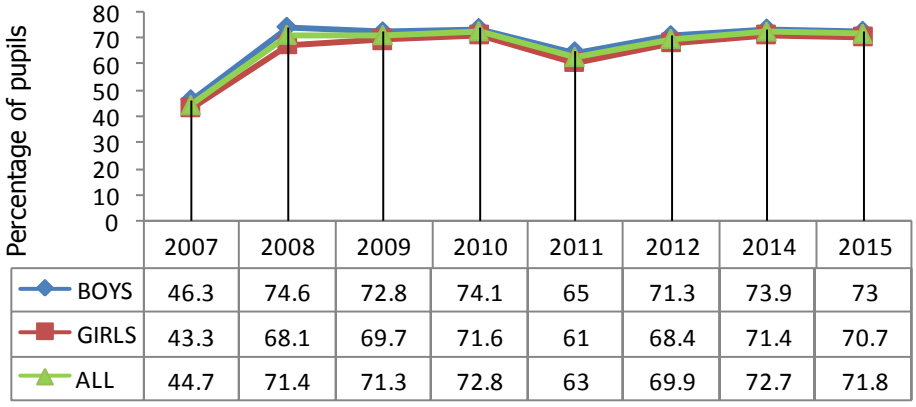


In secondary Mathematics, there are some items that present extreme difficulties – for example, item 4 requires the use of a set symbol to represent that the second of two listed sets is a proper subset of the first. It is easy to make any one of several mistakes. Item 40b, in which there is little difference between full and part marks, requires an understanding of reflection on Cartesian axes.

B.3.3 Time comparisons

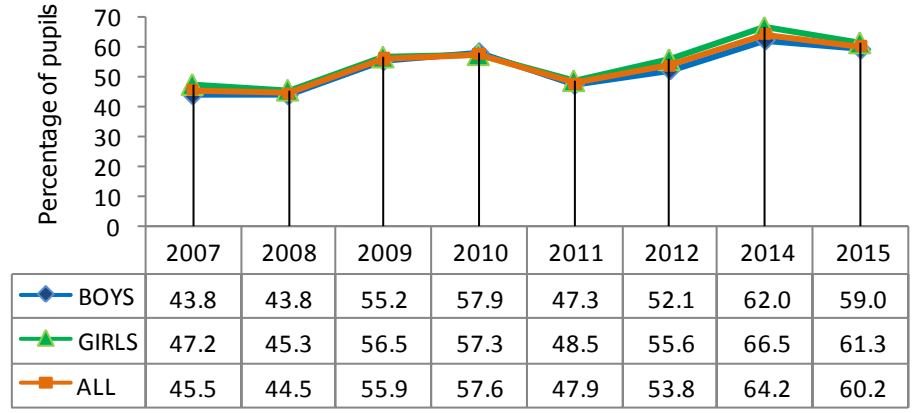
B.3.3.1 Overview of trends from the NAPE reports

FIGURE 3.06: PERCENTAGE OF P 3 PUPILS RATED PROFICIENT IN NUMERACY OVER THE YEARS 2007-2015, BY GENDER



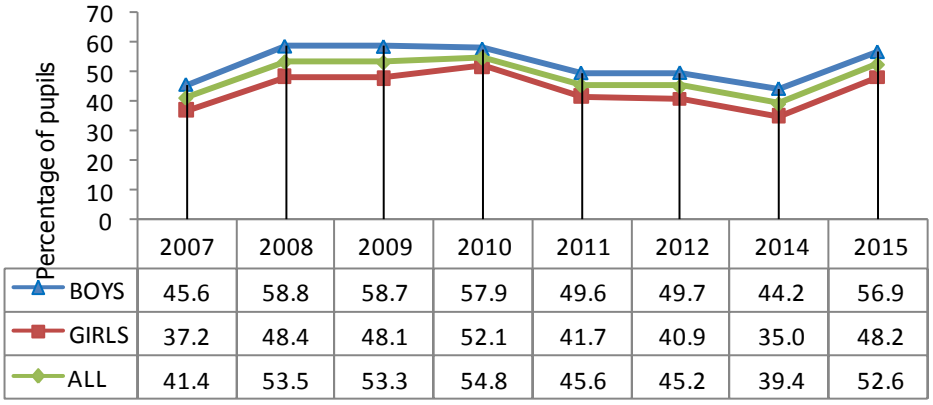
	2007	2008	2009	2010	2011	2012	2014	2015
BOYS	46.3	74.6	72.8	74.1	65	71.3	73.9	73
GIRLS	43.3	68.1	69.7	71.6	61	68.4	71.4	70.7
ALL	44.7	71.4	71.3	72.8	63	69.9	72.7	71.8

FIGURE 4.06: PERCENTAGE OF P 3 PUPILS RATED PROFICIENT IN LITERACY IN ENGLISH OVER THE YEARS 2007 - 2015, BY GENDER



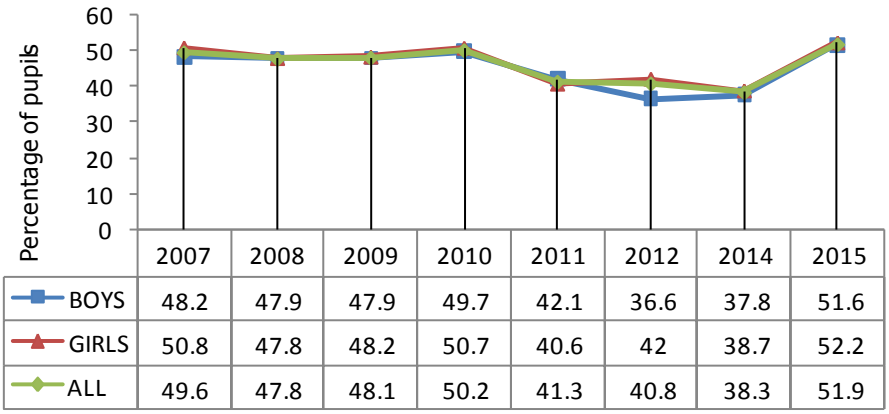
	2007	2008	2009	2010	2011	2012	2014	2015
BOYS	43.8	43.8	55.2	57.9	47.3	52.1	62.0	59.0
GIRLS	47.2	45.3	56.5	57.3	48.5	55.6	66.5	61.3
ALL	45.5	44.5	55.9	57.6	47.9	53.8	64.2	60.2

FIGURE 5.06: PERCENTAGE OF P 6 PUPILS RATED PROFICIENT IN NUMERACY OVER THE YEARS 2007 - 2015, BY GENDER



	2007	2008	2009	2010	2011	2012	2014	2015
BOYS	45.6	58.8	58.7	57.9	49.6	49.7	44.2	56.9
GIRLS	37.2	48.4	48.1	52.1	41.7	40.9	35.0	48.2
ALL	41.4	53.5	53.3	54.8	45.6	45.2	39.4	52.6

FIGURE 6.06: PERCENTAGE OF P 6 PUPILS RATED PROFICIENT IN LITERACY IN ENGLISH OVER THE YEARS 2007 - 2015, BY GENDER



	2007	2008	2009	2010	2011	2012	2014	2015
BOYS	48.2	47.9	47.9	49.7	42.1	36.6	37.8	51.6
GIRLS	50.8	47.8	48.2	50.7	40.6	42	38.7	52.2
ALL	49.6	47.8	48.1	50.2	41.3	40.8	38.3	51.9

Source: NAPE 2016 summary report

Trend data are presented since 2007. Apparently radical changes are shown from one year to another – changes that may be attributable to a change in the test content or difficulty, or to changes in standards setting. The trend data show important variations in the results (+/- 10 points for some years). The standard errors and confidence could be graphically presented rather than being left out or hidden in technical manuals. There are no metadata on the comparability of tests across years in the report. The reader does not know if similar tests were used across years or if there were anchor items.

The test length (number of items) varies across years in primary. The construct measured in primary for literacy varies across years, sometimes including recopying or drawing. Since the indicators have large variation from one year to the other it is likely that the changes observed result from a difference in test difficulty or in sampling procedures, and not from a change in pupils' learning outcomes. The use of anchor items is unknown.

B.4 Conclusion on validity

The objectives of the NAPE are:

1. to determine and monitor the level of achievement of pupils over time;
2. to generate information on what pupils know and can do in different curricular areas;
3. to evaluate the effectiveness of reforms in the education system;
4. to provide information on variables that affect learning achievement;
5. to suggest measures for the improvement of teaching and learning in schools; and
6. to provide data for planning and research.

The uncertainty over the meaning and interpretation of cross-years data and the small sample size for districts present difficulties in using the results to describe changes and differences (objectives 1 and 2) and to evaluate the effectiveness of reforms (objective 3). The nature of the test and the data about schools, pupils and contexts collected with it present difficulties in using the results to achieve objectives 4, 5 and 6.

Last but not least, the tests seem to show construct under-representation in maths: there are not enough items measuring high cognitive abilities, such as complex applications that should be targeted in the syllabus. The maths test does not tell us 'what pupils can do' since it is restricted to lower cognitive levels. The P3 sub-component of the literacy tests administrated every three years is not a valid measure of basic literacy skills. Data show it is not the same construct as the rest of the test and the concepts measured are not clear.

Similar conclusions to ours, based on longer and more comprehensive analysis, including looking at a range of procedural, documentation and reporting issues, were reached in a review of NAPE 2003 over 10 years ago (Clarke 2005). We mention this not as an occasion for fault-finding or reproach but as an example of what the main text describes as one of the key results of Uganda's current assessment system: '*reports by agencies often do not lead to effective action – at central, district and local level*'. The implication of this for the purposes of this road-map is that technical reports are not of themselves a means to enhance Uganda's assessment system. Our proposals for a systems approach are directed towards changing this situation by developing agreement across stakeholders about what should happen, when, how and at what opportunity cost, and then monitoring and reporting on the match of expectations and realities.

B.5 SWOT diagnostic

Strengths

- Maths problems include some related to money and time but could be more diversified in terms of context.
- NAPE seems to be socially accepted (face validity) and seems to be used as a monitoring tool for learning outcomes (by donors in particular).
- Data are used for secondary analysis (World Bank[23] for instance).

[23] http://documents.worldbank.org/curated/en/2013/02/17406414/uganda-improving-learning-uganda-vol-3-3-school-based-management-policy-functionality.

- The procedures allow people with disabilities to take the test[24].
- The S2 tests have good psychometric properties.

Weaknesses

- The tests exhibit some redundancy[25] and dependency in items.
- The repartition of items by domain/strand and cognitive level is not as useful as it might be.

Opportunities

- As one of the oldest national assessment of learning outcomes in Africa (1996), the institution and its staff have a lot of experience.
- This is a long test that would make it possible to measure several competences.
- Some forced choice (multiple choice items) could be used to reduce the burden of correction.

Threats

- In the absence of documentation of standard errors, indicators at district and learning outcome level may be misinterpreted.
- The tests we analysed may provide an overly positive evaluation of primary pupils' achievement, and thus lead to an underestimation of the challenges of teaching and learning in primary.
- The use of constructed response items (rather than multiple choice items) increases costs, possibly without always sufficiently increasing the information that is generated. On the other hand, good multiple choice items are expensive to develop and this may be one of the reasons their use was phased out.

[24] http://www.uis.unesco.org/nada/en/index.php/catalogue/103.
[25] This is a technical term in psychometrics. See http://www.rasch.org/rmt/rmt143a.htm for a brief summary.

Annex C Analysis of PLE and UCE examination papers and commercial practice tests

A review of PLE examinations and a sample of UCE examinations and the relevant syllabuses suggests a range of differences between the priorities stated in curriculum documents and the priorities that seem apparent in assessment practices.

C.1 Overview

The purpose of this annex is to summarise what we learned from a review of examination papers and examples of tests, which appear to dominate classroom assessment – in the sense both of frequent testing (sometimes three times a term) and of providing the sorts of exercises completed in class and marked by teachers. We sought to identify the recurrent and dominant features of *realities*, not intentions or aspirations. Describing such realities is not intended to be and should not be taken as criticism or fault-finding of the work of those involved but as a reflection of the current situation. It is clear, for example, that hundreds of examination questions are developed, administered and marked every year, carefully and conscientiously, in ways that involve many people. What they do and how they do it reflects practices and expectations (by and from teachers, students and parents) that have developed over a long time. We understand that the examinations in which these many questions appear meet official requirements.

Focusing on the recurrent and dominant features means that we are certain to have left out references to important nuances, emergent developments and exceptions. More detailed analysis is likely to feature in subsequent phases of this work to support the enhancement processes described in the main report. This is likely to require technical capacity-building to support the well-developed experience and expertise of those responsible for these examinations.

From the information in Annex A it seems that examinations are mainly focused on their selection function: identifying high achieving students. Our review of examination and commercial test papers supplements this observation: we refer to differences between the priorities stated in curriculum documents and the priorities that seem apparent in assessment practices.

Remarks we make about the relationship of examinations and curriculum documents should be understood in the context of the comment of Penny *et al*. 2008 that curriculum reform – a sector-wide activity – functions 'primarily in the formal sphere and not at the level of the experience of most teachers, pupils and their families, yet it is at this level that national education policies have to be mediated in practice.' Our focus is on these realities and nothing we say should be taken as a negative comment about Uganda's sector-wide policies.

C.2 PLE

For our review we used PLE examination papers for 2013, 2014 and 2015, some examples of commercial tests used in schools at P4 and P7 level, current curriculum documents for P6 and P7, and examiners' reports from 2013.

The PLE examination papers become public documents once the examination has been administered and hence there are no items common in successive years, which makes it impossible to apply conventional methods for aligning standards from year to year. This in turn means that despite equivalence in construct, length and duration the alignment of the standards

underpinning conversion of marks[26] into divisions is not clear enough to allow accurate determination of trends in the PLE results data.

PLE papers from 2013 to 2015 are recognisably similar to each other, with minor variations. There appears to be the traditional see-saw between examiners and those preparing students for the examination: 'Will there be a question on this year's paper like this one from last year or the year before?'. For example, there was no question involving ratios in the 2014 and 2015 Mathematics examination – perhaps one is due in 2016.

Detailed comments on the PLE papers from 2015 appear in Annex A.

A comparison of the PLE English examinations and the P6 and P7 curriculum documents suggest a range of ways in which the curriculum documents are not well connected with the examinations. The curriculum speaks of relevant topics, contextual learning, writing, speaking, reading and listening. The PLE examinations are decontextualised. We do not see a clear connection between PLE P6 and P7 syllabus – relevant topics, contextual learning, writing, speaking, reading, listening and the decontextualised PLE examinations with their strong emphasis on single word answers, formal restructuring of sentences and comprehension items that are limited to finding a matching word or phrase in a passage. Examiners' reports about students' responses are primarily variations on themes such as 'weak candidates didn't read the question', 'didn't understand the passage' and 'gave the wrong answer'.

The P6 and P7 mathematics curriculum speaks of the importance of higher order thinking and creativity – neither are clearly present in the examination papers. While further detail seems mostly to emphasise learning and applying methods (for example, converting numbers from one representation to another) that do play a large part in the examinations, the curriculum also mentions some complex concepts – for example, infinite sets – that do not seem to appear in the examinations.

The divergence between concepts specified in the curriculum and examination questions is illustrated by an example, ostensibly about probability:

'In a car park there are 192 cars. The probability that a car picked at random from the park is made in Japan is 5 / 8. How many cars are not made in Japan?'

While it is conceptually sound to infer an estimate of probability from a frequency, it is appropriate to deduce actual occurring frequency from an estimated probability – probabilities deal with possibilities not actualities (some event that has actually happened has a probability of 1.0 regardless of its probability before the event). In this case, the question is more an application of a fraction – for example, it could have read as follows: 'Exactly 5/8 of the cars were made in Japan. How many were not made in Japan?'.

The examiners' reports in mathematics mostly state candidates' weaknesses in terms of failing to do what the question asked, rather than understanding their reasoning and approach. Advice to teachers often emphasises content, definitions and the need for practice.

For PLE science, the curriculum includes the importance of scientific processes – understanding and using them, including the idea of hypotheses and analysing cause and effects. The

[26] The processes used to convert scores on one year's PLE examination into divisions are not publicly documented. We expect that these processes draw on considerable experience of key staff. We can see from media coverage that the public and teachers clearly expect that divisions are comparable in standard from year to year and there are some perceptions that standards are falling. There are procedures that can be used to demonstrate the nature and extent of such comparability,

examinations focus primarily on recall of learned content. For example, one item in the 2014 paper is 'How is the BCG vaccine different from the polio vaccine in the way they are given to children?'. At this time, news reports indicated that Uganda was moving from oral polio vaccines to injectable polio vaccines. There is a conceptual difference between the BCG and polio vaccines that is not captured by the difference in method of administration. In the same paper, there is an item that asks the student to 'give any one disadvantage of harvesting honey using fire'. It seems unlikely that a response along the lines of 'it melts the honey so it runs away' would receive any credit, in comparison with a recalled response of 'it kills the bees' or 'it pollutes the honey'.

The examiners' reports in science mostly state candidates' weaknesses in terms of failing to do what the question required, whether this be a statement of fact or giving a reason. Advice to teachers often emphasises helping, showing and explaining.

For PLE SST[27], the curriculum describes clearly the value of social studies in preparing students to be informed and participating citizens, with problem-solving skills and understanding of concepts of 'physical, social, political and economic factors which will enable them to cope with multiple perspectives in the dynamic world as well as critical and creative thinking skills'. The more detailed content descriptions, however, emphasise naming and listing. The examination papers similarly place an emphasis on recall, with items often framed as 'state', 'name' or 'give a way in which'.

The examiners' reports in SST mostly state candidates' weaknesses in terms of failing to do what the question required, whether this be a statement of fact, a giving a reason or naming a feature on a map. Advice to teachers often emphasises helping, showing, mentioning and explaining.

We reviewed some examples of commercially available tests used in P4 and P7. These, as might be expected, are essentially items copied from or adapted from items in PLE examinations. Student responses on these tests (which, as discussed in Annex D, are used as a basis for reporting to parents) indicate that they are learning how to make their responses to such questions routine and rapid. Other than the items that require straight recall, students develop (or are guided by their teachers to develop) ways to identify a particular item as one of a kind with which they are familiar, and to apply a well-rehearsed approach (or they continue making no sense of what they are doing).

In the P4 practice examinations we reviewed, some students were clearly writing something in the response area that had no connection with the question or that may have repeated a word or a word fragment from a neighbouring question. For comprehension items – ones that are at P6 level – some students may simply write a word from the passage.

C.3 UCE examinations

There are 49 UCE examinations with results recorded in the sample dataset discussed in Annex A. Almost all of these are taken by very small numbers of candidates.

Syllabuses have been reviewed in recent years and new ones developed – a process that has included consultations with employers that have identified the importance to employers of students' having 'generic skills' and their perception that students who have succeeded in education up to S4 level have **lower** levels of these skills than new recruits in general (Guillermo 2012). It is not surprising, therefore, that we find little evidence that these skills are prominent in current UCE examinations,

[27] We deliberately make no observations on the religious education aspects of the curriculum or the examinations.

We have reviewed examinations in English Language, Literature in English, Physics, Chemistry, History and Biology, with reference to the syllabuses (those we had available at the time of review). These are the syllabuses published by NCDC in 2008, such as the Chemistry Teaching Syllabus Senior 1-4 or the Physics Teaching Syllabus Senior 1-4. We have also reviewed examinations in Mathematics. Examiners' reports from 2013 have been very helpful.

The syllabuses we reviewed vary in the specificity of the details they provide about summative assessment. Our remarks that follow should be understood with this limitation in mind and should not be seen as a comment on the technical adequacy of the examinations.

The UCE examination papers become public documents once the examination has been administered, and hence there are no items in common in successive years, which makes it impossible to apply conventional methods for aligning standards from year to year. This in turn means that despite equivalence in construct, length and duration over time (there have been some changes in some subjects) the alignment of the standards underpinning conversion of marks[28] into divisions is not clear enough to allow accurate determination of trends in the UCE results data.

There is an integrated UCE English syllabus that was published in 2008 (apparently as an interim measure). The relationship between this syllabus and the English Language and the Literature in English examinations is not clearly evident. It appears that past papers serve as a clear guide to what is expected in the next examination. At this stage, as required by the syllabus, the English examinations assess reading and writing macro skills, leaving out speaking and listening.

The examination papers in Literature in English are, given the wide choice of set books, mostly template questions, such as questions of the form 'What lessons do you learn from [the relationship of][what happens to] [characters]?' or of the form 'What problems does [the character] face?'. These two types of question are fundamentally different in principle, though possibly not in practice, since the second type does not require the student to articulate a personal response. Questions of the form 'What makes [this work/poem] interesting?' in principle allow for negative responses ('nothing', 'the front cover') but it appears likely that such personal responses would be unlikely to be well regarded.

Physics has three papers, including a practical paper. The syllabus places an emphasis on concepts, applications, and analysis and synthesis of ideas, application to new situations, *devising* experiments to test hypotheses and so on, although its assessment section appears to be silent as the extent to which these should appear in summative assessment and the syllabus details place greater emphasis on definitions, descriptions and identification.

We understand that the resource requirements of the practical paper may not be met to the same extent for all candidates – with an obvious impact on their potential results.

Setting the multiple choice component presents some challenges since a trial on Ugandan students is not possible. These items are more oriented towards recall than towards concepts. There are, understandably, some wording issues from time to time. For example, one item reads 'When an a.c. is applied to the Y-plates of a C.R.O and the time-base is switched off, the peak voltage of the a.c. is proportional to …', although the correct fact is that the vertical line on the screen (the effect) is proportional to the a.c. voltage (the cause), not the other way round – as is expected by this question. Again, 'Electric field lines always move from positive charge to negative

[28] The processes used to convert scores for one year's PLE examination into divisions are not publicly documented. We expect that these processes draw on the considerable experience of key staff. We can see from media coverage that the public and teachers clearly expect that divisions are comparable in standard from year to year. There are procedures that can be used to demonstrate the nature and extent of such comparability,

charge' could distract some students, since the lines themselves do not move – they point in the direction a positively charged source would move if placed in this field. The extent to which this sort of feature is a distraction for very able students is not known – item analysis would be useful.

The examination papers and examiners' report about the practical examination suggests that there is no clearly evident assessment of what the curriculum refers to as synthesis of ideas, application to new situations and the exercise of evaluative judgement. This last is a feature that is clearly omitted from the list of requirements for the practical examination.

In UCE History, there have been changes in the length and format of the examination since the syllabus was published: a longer examination is now in place, requiring students to answer four questions out of eight rather than three out of eight.

Questions tend to be of the form 'state/describe' (placing a premium on recall), together with, in some questions but not all, 'explain/why'. Consistent with the syllabus, there does not appear to be any notion of problematic, multiple or contested interpretations of history.

The UCE Biology syllabus sets out its emphasis on applications to everyday life and the importance of investigations. These features appear in the examination papers. Many of the questions, however, seem to rely more on correct recognition and use of specialist terminology to make them difficult than on assessing concepts.

The use of 'explain' in the UCE Chemistry syllabus seems to mean something more along the lines of 'describe' (as in, for example, page 51 of the syllabus: 'Explain the pollution effects'), rather than referring to deeper and more complex scientific explanations. The idea of 'explain' is not, however, much emphasised in the examinations. Instead, the examiners' comments show that the students are expected to use chemical reasoning to work out the answers – a complex and multi-faceted process requiring excellent recall of many details, including reactions, methods, chemical properties and calculation methods. This emphasis is reflected in the way that the results in this subject identify a small proportion of high achieving students. Analysis of examinations at the item level using IRT would help to refine our understanding of these issues.

Sometimes, of course, what the examiners expect a question to assess is more than is needed. For example, one question asks 'Which one of the following reacts with cold water to produce hydrogen: aluminium, zinc, copper, calcium?'. While the examiners' report said that this item required recall of reactivity series, it seems that ordinary, real-life experience of cooking utensils, screens and electrical wire identifies the only plausible choice.

UCE Chemistry Paper 1 has all the hallmarks of an examination focused on discrimination: there are 50 questions (often complex and tricky) to complete in 90 minutes, including questions with multiple parts at the end of the paper.

The UCE Mathematics papers present a view of mathematics as involving techniques rather than involving conceptual understanding or practical applications. For example, there is a question ostensibly about the construction of a shed using iron sheets for the roof. The actual answers to the number of sheets needed and the total cost requires a completely unrealistic assumption that the sheets are laid without any overlap (and without any wastage). In these questions, logarithm tables are used to do calculations, vehicles travel at uniform speeds (presumably on straight, well-made roads), money deposits earn simple annual (not daily) interest without compounding and coins are unbiased. This is, of course, the mathematical world, not the real world of today, and a fascinating and important world it is, but it is different from the real world, where the application of mathematics is more complex (and even more interesting). Employers expect recruits to have

mathematical skills that can be and are applied to the real world – being able to get good results on these sorts of mathematics examinations is, experience suggests, often not a sign of being able to use the sorts of skills employers expect but it is a sign of likely success in further study.

Annex D School-based fieldwork

Our fieldwork suggests that everyday realities of Ugandan classrooms and schools present significant challenges in the implementation of changes in classroom assessment practices.

D.1 Introduction

The collection of primary data in school was conducted to reach an understanding of the challenges involved in developing an approach for integrating effective classroom assessment into Uganda's assessment system.

The aim of the school-based fieldwork was to:

- gain an understanding of the types of assessments used by teachers in primary schools;
- establish the extent to which teachers use assessment to inform their teaching;
- identify the challenges teachers face when assessing students; and
- establish why and how (head) teachers use commercial tests.

D.2 Approach

As part of Phase I of the process, primary data collection was undertaken in a small number of purposively sampled schools in order to ensure the road-map responds to the realities of teachers on the ground. The small number of schools and the limited time limits the richness of the data collected in this phase.

It is proposed that data be collected in a larger sample of schools during Phase II of the process to capture the full range of heterogeneity in approaches to assessment across schools in Uganda, and to refine and develop the road-map. This larger collection should have both a quantitative element across a larger number of schools, to collect prevalence data, while using qualitative studies of a smaller number of schools to gain essential insights into, and a deeper understanding of, the factors that influence assessment practices.

The school-based fieldwork uses both confirmatory and explanatory analysis to understand how assessment is undertaken in primary schools, identify challenges teachers face and understand why (head) teachers use commercial tests in the schools selected for the qualitative research.

The qualitative research was designed and implemented by a team consisting of both assessment researchers and government officials representing the government bodies who have a mandate to carry out assessment in Uganda. In order to ensure the interpretations and findings of data resulting from the fieldwork are grounded in the context and structures in which schools function, NCDC, DES, TIET and UNEB were invited to select a representative to participate in: (1) the sampling of schools based on the sampling methodology; (2) development of tools; (3) data collection in schools; and (4) the verification of findings.

D.3 Methods

Qualitative research is often accused of being open to research bias or anecdotal impressions, impossible to reproduce and difficult to generalise (Mays and Pope, 1995). We address these risks through a robust qualitative research design through sampling and analysis. Each of these criticisms is discussed in the relevant sections below.

Methodological rigour in qualitative research is not established through a statistically representative sample, because results cannot be quantified and aggregated in the same way as quantitative data. Nonetheless, as in quantitative research, rigour in qualitative research can be achieved through '*systematic and self-conscious research design, data collection, interpretation and communication*' (Mays and Pope 1995: 110). Moreover, '*as in quantitative research, sampling is a core design issue because the purposeful sample will determine what you learn about*' (Patton, 2015).

As with most qualitative research, the chosen approach to sampling for the qualitative component of the school-based fieldwork is designed to generate responses from small numbers of individuals and groups. For Phase I, schools were sampled through typical and extreme case sampling, so that the small number of schools are most likely to give an idea of the **range** (rather than the distribution or prevalence) of issues and challenges facing schools and teachers (and the Ugandan education system) as they seek to develop effective classroom assessment.

The generalisability of the qualitative research results derives from the extent to which they have some validity in a wider context. This form of sampling allows exploration of how assessment takes place in a typical case, but also in 'good' and 'bad' cases, which helps explain the reasons behind why stakeholders behave in different (or similar) ways across contexts.

However, given the time and resource constraints of the Phase I fieldwork component, the geographical coverage and sampling approaches were extremely limited. Therefore, it is acknowledged that the full range of issues and challenges facing schools and teachers in Uganda may not have been encountered within the schools sampled. Greater generalisability can be achieved through a larger study in Phase II.

The fieldwork component of this research undertook the following sampling:

- eight schools were sampled as sites for the qualitative research;
- four of these schools were based in Uganda's Eastern Region, and key quantitative indicators indicated that the schools achieve reasonably different learning outcomes;
- four of these schools were based in Uganda's Central Region, and key quantitative indicators indicated that the schools achieve reasonably different learning outcomes;
- two of these were 'good case' schools for each region (Eastern and Central), and key quantitative indicators indicated that the schools have higher learning outcomes, compared to other Ugandan schools; and
- two of these were 'bad case' schools for each region (Eastern and Central), and key quantitative indicators indicated that the schools have lower learning outcomes, compared to other Ugandan schools.

The eight selected schools represent a variety of communities with reference to school size, rural/urban location, school performance, public/private provision, religious and ethnic diversity and socioeconomic status.

D.3.1 Qualitative instruments

During the fieldwork period classroom observations were administered and key informant interviews (KIIs) were conducted with teachers, parents, and head teachers. The objective was to gather data on community, teacher, parent and head teacher perceptions and practices. The tools for the KIIs and classroom observations were developed in collaboration with government officials in order to capture information on the core areas to probe.

D.3.2 Analytical approach

The analysis uses applied thematic analysis, primarily to confirm a set of hypotheses – also known as classic content analysis (Krippendorff, 2004). The selected principal approach – 'confirmatory analysis' – aims to confirm a set of pre-existing hypotheses. This is in contrast to exploratory analysis, which derives hypotheses from the data collected (Guest *et al*., 2012). Exploratory analysis is used as a secondary analytical technique, to ensure that the qualitative component is responsive to new information. Confirmatory analysis is then used to validate new information.

Applied thematic analysis requires researchers to interpret data and does not rely on counting words or phrases, but rather identifies and describes implicit and explicit ideas that are organised into themes.

The following matrix outlines the areas of exploration, the pre-existing hypotheses that require testing, the core areas probed and the sources of information.

Table 7: Matrix to inform and analyse the qualitative data

Area of exploration	Hypotheses	Core areas to probe	Source of information
The different types of assessment used by teachers in primary schools	Many schools implement extensive and intensive testing, and they often use scarce resources to purchase commercial tests	How often students are assessed and which forms of assessment are used. How much moneys schools use to purchase commercial tests. Sources of resources used to purchase commercial tests	Head teacher KII; teacher KII; parent KII
The extent to which teachers use assessment to inform their teaching	Many schools do not systematically plan and report on student achievement through classroom assessment designed, implemented and recorded by teachers. Teaching is informed by patterns defined by past exam papers consisting mainly of items requiring recall or definitions, or of how to do 'rehearsed' items	Teacher perceptions of the purposes of assessment. Teacher practices in assessment. How teachers use assessment to inform their teaching	Teacher planning documentation; student achievement reports; teacher KII; head Teacher KII
Challenges teachers face when assessing students	Teachers face pressure from parents and schools to get the highest possible marks in PLE examinations, which results in a narrowing of the curriculum in classrooms. Many teachers focus their teaching towards students' maximising their marks in examinations through 'coaching' or 'cramming'. 'Coaching' and 'cramming' inappropriately narrows the focus of learning in classrooms. The full range of competences specified in the syllabus is not sufficiently included in teaching and assessment in schools. The knowledge, skills and competences students need for success in their future education, life and career are not sufficiently included in teaching and assessment in schools, as would be expected by the broader community. Teachers do not receive professional development to impart the knowledge and skills required to use classroom assessment effectively	Head teacher decision-making regarding how, when and what to assess. Teacher decision-making regarding how, when and what to assess. Teacher decision-making regarding how and what to teach. Parental perceptions of what is assessed in schools. Parental perceptions of what should be assessed in schools. Parental perceptions of what knowledge, skills and competences students need for success in their future education, life and career. Teacher self-efficacy in classroom assessment	Teacher KIIs; head Teacher KIIs; parental KIIs; classroom observations
The use of commercial tests in schools	School accountability is predominantly, in practice, expressed through examination marks – therefore schools focus on narrow and rote test-taking processes (using commercial tests) in order to prepare students, as it is in their interest to support a narrow focus on teaching and learning. Parents are conflicted between the immediate needs of their children to pass high-stakes examinations and their	How head teachers assess school performance. How parents assess school performance. How teachers assess school performance. The role of private tutoring in the testing process. Who administers commercial assessments? Who pays for them?	Head teacher KIIs; teacher KIIs; parental KIIs

	desire for their children to have a fully rounded education. Therefore, while parents may prefer that their children be exposed to broader teaching and learning experiences, they invest in or support the use of commercial tests to prepare students for examinations to secure their immediate futures.		

D.3.3 Timing

Fieldwork preparation was undertaken in Kampala in the week of 14 March 2016. During this time two qualitative researchers from the Examinations and Assessment Road-map team worked with government officials to select schools and develop instruments based on Table 7: Matrix to inform and analyse the qualitative data. One qualitative researcher then conducted the fieldwork together with two government officials in the Eastern Region and then also in the Central Region.

Field visits took place during the period 21–23 March and 4–6 April.

D.4 Schools visited

In order to retain confidentiality and anonymity for the research participants, the schools have been given pseudonyms. A brief description of the key characteristics of each school is provided below:

- **Eastern high performing private school** is in a peri-urban location on the outskirts of a district town. It is a relatively small private school, with an average of 35 pupils per class.
- **Eastern high performing government** school is regarded as a relatively high performing government school in its district. It is located within a rural setting, has just over 50 students per class, and is managed by the Catholic Church.
- **Eastern poorly performing private school** is regarded as a poorly performing private school. Located in a district town, it has approximately 175 students.
- **Eastern poorly performing government school** is regarded as a low performing government school. It is located about 15km from the district town, in a rural location, and in a seemingly disadvantaged area.
- **Central (Masaka) high performing private school** is an urban school, on the edge of the district town. It has over 1,000 students on a very small site.
- **Central (Ssembabule) high performing private** is regarded as a relatively high performing private school in the district town. It has around 35 students per class.
- **Central (Masaka) poorly performing government** is located in a rural area, and has around 600 students in total.
- **Central (Ssembabule) poorly performing government school** is located in a rural, very disadvantaged area. At the time of fieldwork, it was planting season, which meant that up to half the students were absent.

D.5 Findings

The findings of the small-scale study are organised around three themes, developed from those listed in Table 7. Within each theme one or more hypothesis is tested.

D.5.1 Types of assessment used by teachers in primary schools

The types of assessment used by teachers in primary schools was explored, with specific reference to seeking to understand how often students are assessed and which forms of assessment are used, how much money schools use to purchase commercial tests and the sources of resources used to purchase commercial tests.

Across the range of schools visited it appeared that all schools implemented extensive and intensive testing. Typically, this involves homework exercises after each lesson or a weekly written test; and a more formal exam-like test at the beginning, middle and end of term. The beginning, middle and end of term tests typically cover all subjects and take at least a whole day, during which time no lessons take place. A larger investigation would identify the extent to which this is a common practice in both private and government schools.

It was reported to the researchers that it was typical across schools for weekly or daily exercises to be written by the teachers themselves, with the exception of one school where CIPRO lesson plans were used, which include an assessment task[29]. Within the classroom observations it was observed that formative assessment tasks were based on lower order skills, such as repeating information from the lesson undertaken. In some schools the assessment task simply required copying from the board.

The data indicate that teachers are aware of the importance of good questioning in class. However, almost all teachers observed asked only closed questions. In many cases the whole class responded in unison, preventing the teacher from identifying which children had understood the lesson. For example, in a P2 class, students were encouraged to read 'hospital' from the blackboard. One student could do so, and the others chanted along soon after. The teacher did not follow this up and help the students who could not read the word.

The data indicate that in all schools within the sample (see the listing in section D4), scarce resources are used to buy commercial tests to some extent.

Using resources to buy commercial tests varied:

- some schools bought three sets of tests per term for all students (Eastern high performing government school, Eastern poor performing government, Central (Masaka) high performing private, Central (Masaka) low performing government);
- some schools bought tests just at the end of term (Central (Ssembabule) low performing government and Central (Ssembabule) high performing private school); and
- some schools bought tests as and when funds were available (Eastern low performing government school).

The Eastern high performing government and the Central (Ssembabule) high performing private schools fully grasped the weaknesses of commercial tests, and administered them occasionally as an external check, but preferred to encourage teachers to create their own. In the Eastern high performing private school, teachers could provide evidence of the tests they had written on this basis. In the Eastern low performing government school, the head teacher suggested that the tests were sometimes written internally and sometimes bought commercially, however there was little documentary evidence of internally produced assessments.

Given that commercial tests typically cost 750–1,000 shillings per student per administration, the use of commercial tests can cost up to 1 million shillings per term for a school of 300 students, if

[29] CIPRO is a commercial test provider.

they are used at beginning, middle and end of term. CIPRO was the most commonly cited commercial test, followed by PRIME.

D.5.2 The extent to which teachers use assessment to inform their teaching

The extent to which teachers use assessment to inform their teaching was explored within the case study schools. Discussion with teachers sought to understand teacher perceptions of the purposes of assessment, teacher practices in assessment and how teachers use assessment to inform their teaching.

Interviewed teachers did not explicitly inform the researchers that their key focus was on preparing students for PLE. Rather, teachers described themselves as assessing knowledge, understanding and the application of knowledge and skills. However, in practice the majority of the assessment observed focus on surface knowledge, requiring pupils to recall or copy information from the board. It was unclear if teachers are encouraged to assess higher order skills but do not know how to do so in practice, or if teachers are purposefully teaching to a specific type of assessment item.

The data indicate that the majority of teachers in the case study schools do plan, record and report on student achievement, although this reporting and recording may take a very limited form. This inference is supported in that lesson plans cited tended to include a description of an assessment task, around half of the teachers could produce a log of assessment achievement and exercise books were often marked in order to show parents the results achieved by their children.

However, these practices were found to focus essentially on tasks that assess recall or copying skills, and there was little evidence that assessment logs were used to inform teaching and learning activities. Corrections in exercise books were predominantly ticks and crosses, and assignment to a remedial class was the most common teacher response to low marks. In most schools there was little evidence that lesson subject matter, the teacher's pedagogical approach or the targeting of content for specific students was adapted as a result of assessment results. A notable exception was Eastern (Masaka) high performing private school, where achievement records were extensively analysed and a second member of a 'team teaching' group was responsible for providing specific feedback to students.

D.5.3 Challenges teachers face when assessing students

The challenges that teachers face when assessing students, ones that are considerably greater in large classes, were explored through probing head teachers and teachers regarding their decision-making about how, when and what to assess and teacher self-efficacy in classroom assessment. In addition, the parental perceptions of what is assessed in schools and what should be assessed in schools was explored. This included parents' perceptions of what knowledge, skills and competences students need for success in their future education, life and career.

The data from the case study schools suggest that parents often do apply pressure to teachers to show good assessment results, including at PLE level. Parents look carefully at PLE results, especially the proportion of students with the highest possible results, and this is clearly factored into decision-making, for example about which school to send their child to. However, parents also have much broader perceptions of what constitutes a good school. Other cited factors included a safe and clean compound, the provision of lunch or their perception of teachers' professionalism. Head teachers recognise the importance that parents attach to PLE results, and are motivated accordingly. However, when asked how they judge their own success or how their superiors judge their school, head teachers also regularly cited enrolment numbers, school infrastructure, and issues around teacher quality, such as regular lesson planning.

In addition, the data suggest that the application of parental pressure on teachers to show good assessment results is also pronounced in terms of ongoing assessments. Within the schools it was often a perception amongst parents that assessments and tests are important aspects of good teaching. Some parents therefore felt that a good school provides more assessments or tests than a weak school. Parents often looked at the results from term-based commercial tests, and some participants stated that the schools responded by teaching to those assessments. This was particularly the case in the Central low performing government (Masaka) school, where a head teacher said that she administered an inappropriate formal test to the pre-primary class because parents demanded it. The assessment was administered in English, a language to which the children had little exposure. The head teacher stated that other schools (particularly private schools) followed this practice, and if the school did not test the young children, parents would move their children to a nearby private school. It appears that parental pressure in regard to obtaining good assessment results reflects, at least in part, real confidence in UNEB and its assessment processes.

The focus of learning in most of the classrooms observed was shallow. Coaching for PLE appeared to be the principal concern, as all stakeholders acknowledged the importance of doing well in these exams. However, it was unclear if the shallow teaching and learning observed was being driven by the PLE exam content, or by a lack of teacher skills to teach, and/or confidence in teaching, a deeper set of knowledge and skills. The interviews indicated that the majority of teachers aspired to teach the skills emphasised in curriculum documents, such as 'creative thinking' or 'independent investigation'; however, this approach was not evident in the actual lessons observed. Teachers did not seem to recognise their approach as 'cramming' or teaching shallow skillsets when this was the practice observed, suggesting a lack of teacher competence to move away from these approaches.

The cycle of three 'exam-style' assessments per term (beginning, middle and end of term tests) was found to reduce learning time in the case study schools. Students were often encouraged to prepare for the assessments in after-school homework classes, for example. In P7 some schools administered weekly tests (exam-style, as opposed to classroom exercises) to prepare for PLE, and this is likely to result in surface teaching practices and a narrowing of the curriculum.

Almost all teachers interviewed said that they felt comfortable with all aspects of assessment. However, very few of the assessments observed were of a high quality. This may suggest that teachers' confidence in their own assessment skills is misplaced, since there is a general lack of knowledge about how to set assessments that go beyond recall and copying tasks. There were a small number of exceptions observed in high performing private schools, in which teachers had written their own assessment questions. These items encouraged students to apply and develop their skills and allowed teachers to differentiate learning needs.

Schemes of work, based on the curriculum, were evident in almost all schools. Teachers were often judged by their head teacher partly on the extent to which they covered the curriculum. Most teachers had lesson plans in place to ensure they covered the curriculum. However, some teachers reported a challenge in covering all the content. As this is a very large aspect to cover in a small-scale study, it is suggested that further research be conducted to investigate the nexus between curriculum coverage and assessment practices in Ugandan classrooms. For example, there may be too much content to cover in the time available, especially after time is given to frequent testing. Again, it may be that some teachers need to use the commercial tests to identify the key aspects of the curriculum they should teach: in this sense, the commercial tests are meeting a need – 'filling a gap' in teacher knowledge and skills.

As stated previously, the Central (Ssembabule) low performing government school is situated within a disadvantaged community. In this school parents were unaware of what is taught and assessed in the school and there was a general lack of engagement between the school and the community. However, in all of the other case study schools parents used the results of tests to make judgements about the quality of the school and the teaching.

Parental perceptions regarding what is assessed in schools was rarely specific. Overwhelmingly, parents stated that children should be tested on what is taught in class, but parents were generally quite unaware of the content of classroom lessons or the syllabus. When probed about what knowledge, skills and competences students need for success in their future, parents stated that children need to be able to help their parents, gain wisdom, and read and write. Some parents also made reference to the importance of children leaving school as good citizens or good members of the community. Several parents cited good co-curricular activities as an important sign of a successful school, or described their vision for the school in terms of producing well rounded school leavers.

Data suggest that parents do seem to want their children to achieve both academic success and a broad education, and do not seem to perceive a conflict between these two goals. Commercial tests are perceived to provide a regular, intelligible and clear way of evaluating school performance, which parents tend to value, with the majority of parents interviewed indicating that they pay for commercial tests when required. There was little suggestion that parents perceived the use of commercial tests as a long-term strategy to prepare for PLE exam questions.

D.6 Conclusions and next steps

The data suggest that many schools implement extensive and intensive testing to varying extents and that often this involves using scarce resources to purchase commercial tests. Many schools do systematically plan and report (if sometimes in a very limited way) on student achievement through classroom assessment but assessments often focus on recall or copying skills. The results of both informal and formal assessments do not tend to impact sufficiently on teaching and learning in classrooms.

The majority of teachers aspired to teach skills such as 'creative thinking' or 'independent investigation', but these were not evident in the actual lessons. Teachers would not recognise the characterisation of their work as 'cramming' or 'teaching a narrow curriculum', even when this was the practice observed. The extent to which narrow assessment is driven by past exam papers is unclear. Low levels of teacher competency in assessment and teaching and learning could be driving 'surface' teaching, rather than (or as well as) previous exam papers. It is proposed that further research be undertaken in Phase II to investigate the reasons for lower order assessment tasks being used. Similarly, teacher responses in these case study schools indicate that teachers have often grasped aspects of the theory of assessment, but continuous professional development in practical assessment skills is both desired and needed. It is proposed that the larger scale study investigate teacher development needs in the area of assessment, to further inform the road-map. Such a study should also look carefully at the different needs of early primary school, especially in the context of the interaction of assessment and mother tongue teaching and the requirements of effective assessment for students with limited literacy,

Stakeholder perceptions regarding the knowledge, skills and competences students need for success in their future education, life and career requires further investigation. Parents were only able to cite very broad competences, including 'helping their parents' and 'good citizenship', however it would be beneficial to explore this further. The concept of 'good citizenship' is globally recognised. However, what it means to be a 'good citizen' is highly contextual and therefore the

meaning of good citizenship within the Ugandan context, and the extent to which this is reflected in Ugandan schooling, is worthy of further exploration.

There does appear to be a relationship between parental pressure to undertake regular assessment, the need for students to achieve high PLE marks, and the high prevalence and weight given to narrowly focused commercial assessments. However, it was not clear how these factors interact. It is proposed that this relationship be explored further in order to identify which factors are drivers, as opposed to consequences.

Several emerging themes indicate that a systems approach is required in order to enact change. For example, the variety of reasons for using commercial tests suggests that these assessments fill important functions in the absence of other options. For example, tests provide teachers who may be unable to develop their own assessments with a pragmatic solution, provide parents with tangible evidence that activities are taking place within lessons, and provide information on school quality and student performance – often in the absence of any other external benchmarks. Simply stating that schools should not use commercial tests is unlikely to change school practices.

Similarly, gaining a deeper understanding of the relationships between particular drivers and unhelpful assessment practices within schools in Phase II of the development of the road-map will allow for the identification of the points of entry available to facilitate change in stakeholder behaviour.

D.7 References

Mays, N. and Pope, C. (1995). Rigour and Qualitative Research, *British Medical Journal* 311(6997): 109–112.

Patton, M.Q. (2015) *Qualitative Research and evaluation methods: Integrating theory and practice*. Fourth Edition. Sage

Guest, G., MacQueen, K. and Namey, E. (2012). *Applied Thematic Analysis*. Sage Publications

Annex E Reports of public discussions of the purposes, products and outcomes of Uganda's assessment system

Reports in newspapers and on-line discussions give us a sense of the public discussion about education and schools.

A 'good' school is commonly seen as one where pupils get high scores in examinations.

There seems to be little public discussion of employer views about the content, standards and results of education, about the role of the community in shaping education and about the aims, goals, purposes and outcomes of education.

The main report describes Uganda's assessment system and identifies that the immediate results of that system appear as a result of what happens in three principal settings: classrooms, public discussion and interactions between and among agencies. The key result of what happens in the public discussion setting is the shaping of the political environment within which agencies – including schools – do their work. The purpose of this annex is to identify the main features of the public discussion about the purposes, products and outcomes of Uganda's assessment system. The data sources for this aspect of the study are primarily newspaper and media reports, illuminated by what we heard from some parents and some of those working in education in a range of roles.

The following account is necessarily interpretive. It does not make any judgement about the accuracy or objectivity of the reports studied, since its purpose is to identify recurrent features (and features that appear conspicuous by their virtual absence) in the public discourse about the purposes, products and outcomes of Uganda's assessment system. There will necessarily be some political aspects underpinning matters and people who are prominent in the news. This account reaches no understanding and takes no views on these issues.

Online articles (newspaper accounts, television news accounts and videos) and attached comments were found using Google search terms with the prefix *Uganda* and [newsmedia[30] name – e.g. *New Vision* or *NBC*] and some combination of three or four of the following terms: *education*, *school*, *PLE*, *UCE Uwezo*, *NAPE*, *standard*, *employer*, *parent.* There is an unavoidable bias in using online resources – a bias that, while voices describing rural situations are still heard, is likely to give more prominence to discourse that is urban and more affluent. Items found in such searches were reviewed if they appeared to be relevant to one or more aspects of Uganda's assessment system (as this term is understood in this report). Recurrent themes were identified in terms of ideas and issues that recurred across time and settings. This focus means we leave out the central topic of news reports based on policy announcements by central agencies or reports of international studies or conferences and specific stories such as a halt on curriculum implementation or examination marking.

The key recurrent themes about the assessment system's purposes, products and outcomes across these sources were:

[30] Some reference to possible source (e.g. New Vision) is necessary to ensure that the many official reports and academic studies in this area do not dominate the Google listings.

1. the idea that a 'good' school is one whose students get 'good' results, get into a high prestige destination (school or university), or get a publicly funded place at university;

2. the value to parents and the community of ranking schools in terms of measures reflecting the proportion of students getting 'good' results (while there are sometimes comments on this as lacking statistical sophistication, these comments take the underlying principle for granted);

3. the interaction of the choice of school attended by students and parental perceptions of social status;

4. differences in the management and performance of public and private schools, and the impact on the 'fairness' of competition between schools;

5. an unspoken assumption that standards implied by the way examination results are reported (for example, PLE division 1) do not change over time;

6. the idea that the role of teaching is to 'coach' students to pass examinations, using past papers as a guide (leading to concerns about fairness if questions are different from expectations);

7. resources, fees and the impact on parents in poverty;

8. examination malpractice of various kinds, at various levels;

9. the size and significance of rural/urban differences in education practices, resources and results; and

10. the idea that levels of education have fallen over time and teachers' status has declined.

The apparent 'silences' in the reports we found have perhaps as great a potential implication for understanding the overall assessment system as these ten themes.

First, there is a pervasive taken-for-granted assumption implicit in much public discourse about 'good' results (with few exceptions) that participation patterns have no impact on the proportion of students with various levels of attainment. This assumption, which is, of course, not generally valid, underpins public discourse about the success of schools, and the meaning of changes in the distributions of PLE and UCE results. Thus, for example, the expansion in access to primary and secondary education should be *expected* to affect the distribution of results, since it is highly likely that the additional participation will involve students from less advantaged backgrounds. Again, the proportion of students at a school getting PLE division 1 results will reflect not only the quality of the school as a whole, and the quality of the teaching and the school's resources, but it will also reflect the quality of the students' family support and their individual ability, attitudes and aspirations, and so on.

Secondly, outside vocational education, there is little indication in these sources of employers' views about education content, standards and results. Comments from parents appear to focus more on primary and secondary education as facilitating access to further stages of education than as developing skills required for success in the 21st century. Employers, it seems, are not adding their views to the public discourse.

Thirdly, there does not appear to be significant recurrent discussion in the community (as opposed to among professionals in the education sector) about the aims, goals, purposes and outcomes of

education, and the community's role in shaping and improving education, schools and outcomes for students.

The change-management for any proposed changes in Uganda's assessment system must consider the implications of the themes and silences described in this annex.

Appendix – Tables and graphs

(See separate attachment called 'Roadmap tables and graphs')