Making educational judgements

Reflections on judging standards of intended and examined curricula

October 2007

PUBLISHED BY

UMALUSI COUNCIL FOR QUALITY ASSURANCE IN GENERAL AND FURTHER EDUCATION AND TRAINING

37 General Van Ryneveld Street
Persequor Technopark
Pretoria, South Africa

Telephone: +27 (0)12 349 1510
Fax: +27 (0)12 349 1511
Email: info@umalusi.org.za

www.umalusi.org.za

ISBN 978-0-620-40310-8

Layout & production
CJ Graphics

Contents

Acknowledgements

This report was written by Matseleng Stephanie Allais.

Jeanne Gamble, Elizabeth Burroughs, and Heidi Bolton were critical readers, and provided valuable advice and suggestions. Pamela Apps edited the report.

The analyses of the different subjects were carried out by the following researchers, and their contribution is thankfully acknowledged:

English: Ana Ferreira, Natasha Distiller, Melanie Grobler, Harriet Davis, and Kay Jaffer.
Biology: Alice Barlow-Zambodla, Heather Blanckensee, Anna Crowe, Edith Dempster, and Michael Mavovana.
Science: Dale Gundry, Makhapa Makhafola, Sharon Grussendorff, Judith Mokhele, Harish Karsan, and Andre van der Hoven.
Mathematics: Hendrik de Kock, Alison Kitto, Themba Mthethwa, Willy Mwakapenda, and Poobhalan Pillay.

The original research was managed by Shireen Badat with assistance from Ramadimetja Phaladi.

Umalusi's Research Forum provided guidance at various points. Members of the research forum over the period during the research included:

Linda Chisholm, Human Sciences Research Council
Magdalena Fourie, University of Stellenbosch
Zelda Groener, University of the Western Cape
Anil Kanjee, Human Sciences Research Council
Peliwe Lolwana, Umalusi
Jeanne Gamble, University of Cape Town
Khulekani Mathe, Procurement Dynamics
Relebohile Moletsane, University of KwaZulu-Natal
Wally Morrow
George Moyo, University of Fort Hare
Joe Muller, University of Cape Town
Prem Naidoo, Council on Higher Education
Herseela Narsee, Department of Education
Ben Parker, South African Qualifications Authority
Yael Shalem, University of the Witwatersrand
Crain Soudien, University of Cape Town
Nick Taylor, Joint Education Trust
Servaas van der Berg, University of Stellenbosch
Nan Yeld, University of Cape Town.

Introduction 1

'Standards' have been a source of ongoing concern for the public as well as for the government and educational providers. In international studies our learners score badly, not only relative to students from wealthier countries, but also to students from poorer countries, both in Africa and elsewhere (Howie 2001; Moloi and Strauss 2005). Contested curriculum reform has also fuelled public debate about the quality of education in the country.

Umalusi has a statutory mandate to monitor the adequacy and suitability of qualifications and curricula and to moderate examinations at both primary and secondary levels. While historically Umalusi's focus (and the focus of its predecessors) has been on examinations, it has become increasingly clear that focusing only on outputs and outcomes is inadequate, and that part of monitoring standards must involve the scrutiny of inputs. Documents representing the intended curriculum are central in this regard. In this context, Umalusi believes that part of its role in monitoring standards must be to evaluate the adequacy of curriculum statements or syllabuses.

Umalusi is now, therefore, moving into a broader role in relation to its mandated competence of judging standards. This includes evaluating the intended curriculum—in other words, the official documents such as syllabuses or curriculum statements that prescribe what should be taught and tested. Umalusi's work is primarily at the level of the intended and the examined curriculum. We do not deal directly with broader aspects of curriculum, such as what actually occurs in the classroom (sometimes called the 'enacted curriculum'). The closest we get to the classroom is the monitoring of institutional quality, which is not the focus of this report.

With regard to examinations, there are various ways in which Umalusi monitors, maintains, and attempts to improve standards. This report only examines one aspect of Umalusi's work in this regard—making judgements about the standards of examination papers. The moderation of question papers is a well developed function at Umalusi and has been undertaken for many years. The criteria currently used by our moderators are attached as Appendix One.

Other ways in which Umalusi monitors standards of examinations include the statistical moderation of results. We are also currently engaged in research to improve our judgements based on learner performance, using, for example, item response theory.

The moderation of question papers remains, however, an important part of the system, and as such, Umalusi needs to ensure that its systems and tools for moderating question papers are as good as possible.

This report reflects the types of tools and processes used in three research projects

which Umalusi has conducted in order to reach judgements about intended and examined curricula, and makes suggestions for Umalusi's systems. This work is part of Umalusi's ongoing attempts to improve the standard of examinations and curricula.

Because evaluations of intended curricula a new area of work for Umalusi, the report begins with a consideration of international literature about what makes a good intended curriculum. This is followed by discussion about ideas about the intended curriculum in South Africa.

The third section of the report provides a description of the three research projects under discussion. Section four discusses what has been learnt from the three projects with regards to processes and systems for evaluating the intended and examined curricula. Section five reflects on what was learnt in the three projects about tools and categories for evaluating intended curricula, and section six similarly reflects on tools and categories for evaluating examinations. The final section considers the way forward for Umalusi.

Thinking about the intended curriculum 2

Standards, curriculum and assessment: International perspectives

Internationally there is remarkable consensus on what primary and secondary education should achieve. According to a Japanese study which examined syllabuses of eighteen countries primarily in Australasia and Europe, the most common goals for the compulsory section of education systems are a basic foundation of knowledge; development of the child's intellectual, emotional, spiritual, and physical potential; and critical thinking and problem solving skills (National Institute for Educational Research 2003). Similarly, Australian research argues that a world class curriculum is characterized by "equity and inclusiveness, the encouragement of innovation and creativity, clarity and focus in content specification and assessment for learning" (Victorian Curriculum and Assessment Authority 2004, p. 3). Even the Japanese, who consistently score highly in most international performance evaluations, are lamenting the need for their children to learn "the ability to identify problem areas for themselves, to learn, think, make judgements and act independently and to be more adept at problem-solving" (Central Council for Education, 1999, p. 18 in Green 2000, p. 425).

Much less resolved, however, is how to design a curriculum that achieves these objectives. Curriculum content has always been contested and there is also much debate about the prescription of the intended curriculum: what should be prescribed, by whom, and in what forms. Most countries have curriculum agencies, bureaus, institutes, centres, or branches of education ministries. Most also have formal or informal consultation with stakeholders. China, Fiji, France, Germany (Bavaria), Indonesia, Japan, Korea, Lao People's Democratic Republic, Malaysia, the Philippines, Sri Lanka, Thailand, and Uzbekistan all report highly centralized curriculum development processes (National Institute for Educational Research 2003). In some countries, there is a varying possibility for local authorities, schools, and teachers to influence curriculum development at the implementation level. For example, local content is allowed in Indonesia (20%), Lao People's Democratic Republic (10%), and Vietnam (15%). In other countries such as Australia and New Zealand, teachers develop their own content within centrally developed curriculum frameworks (National Institute for Educational Research 2003). In the United States, the development of curriculum is totally devolved (although this is more complex than it appears, with some prescriptions at national, state and district levels), while in Germany there is a degree of centralization and a degree of authority at a state (provincial) level (Stevenson and Nerison-Low 2003).

Donnelly (2005) distinguishes between outcomes-based, syllabus-based, and

standards-based systems as ways of manifesting the intended curriculum. A syllabus model, he argues, typically provides detail about what students should be taught at the start of the year, based on established disciplines or categories of knowledge. Specific year levels are identified. Numbers of hours to be spent on topics are often stipulated. A syllabus model is typically associated with summative assessment in the form of high stakes tests.

Outcomes-based models tend to focus on what students should achieve or be able to do by the end of the educational process. They tend to take a multidisciplinary approach, with an emphasis on attitudes, dispositions, and feelings. They suggest stages which incorporate a number of year levels. Numbers of hours to be spent on topics are usually not stipulated. An outcomes-based model tends to take a 'developmental' approach to assessment with a focus on criterion-referenced assessment.

A standards-based model tends to identify what students should know and be able to do at the end of a set time. While this may be based on established disciplines/ categories of knowledge (Schmidt, Wang and McKnight 2005) these are often stipulated in a 'laundry-list' style, as opposed to being presented as a coherent whole, with meaningful relationships between topics, and a sense of the overall discipline. Specific year levels are specified, but numbers of hours to be spent on topics are usually not stipulated. A standards-based model tends to expect that essential knowledge, understanding, and skills are mastered at each year level and are tested through summative assessment. However, Schmidt, Wang, and McKnight (2005) argue that the United States curriculum model, usually regarded as the best example of a standards-based model, is more complex than it appears, because of the long tradition of shared responsibility in curriculum decision-making and a complex decentralized arrangement for schooling and curriculum development.

Donnelly argues that both syllabus and standards-based models tend to be taught with a greater use of direct instruction and explicit teaching, with an emphasis on teacher directed whole-class teaching. Outcomes-based education tends to focus on a constructivist approach to learning, with the teacher seen as a facilitator to a student-centred approach to teaching and learning.

Donnelly argues that countries which have consistently achieved the best results in international achievement tests (including Japan, the Republic of Korea, Singapore, the Netherlands, and Hong Kong) tend to be those with strong syllabus models:

> In countries like Japan and Singapore, syllabus documents leave no doubt about what needs to be taught at each year level, and what children are expected to understand and be able to do. Teachers are free to experiment with sequencing. But, in the main, teachers are expected to devote their energies to improving teaching and hence the quality of student learning. This task is for all teachers in a school - from the newest teacher to the school principal. Mentoring programs are provided for young teachers. Collaborative "lesson study" programs are conducted in each primary school where all teachers work together to refine lessons and to foster high quality learning within and across subject boundaries.
>
> (Donnelly 2005, pp. 8-9)

He goes on to argue that

> One of the acknowledged strengths of a syllabus approach to curriculum development is that each school, and each teacher, does not have to reinvent the wheel by having to design his or her own syllabuses. In Japan, the Republic of Korea and Singapore more time and

> resources, thus, are spent on strengthening lesson preparation and classroom teaching techniques.
>
> (Donnelly 2005, pp. 8-9)

However, as Schmidt, Wang, and McKnight (2005) point out—the issue is not only about content—other aspects, such as presentation of the intended curriculum as a coherent whole, are also important. Donnelly (2005) argues, for example, that it is not only a problem if descriptors of curriculum standards lack academic substance and rigour, but it is equally problematic if there are so many and they are presented in such detail, that teachers are in danger of being overwhelmed. He goes on to argue that this is often a feature of outcomes-based education, because outcome statements are illustrated with multiple indicators and examples that simply add to the checklist mentality and bureaucratic workload associated with implementation. Referring particularly to Australia he comments as follows:

> Australia's adoption of OBE leads to outcome statements that are generally vague, imprecise and lacking in academic content. Many of the Australian curriculum documents seek to remedy the problem of vague outcome statements by listing examples and indicators. Of concern, when compared to a syllabus approach, is that the practice overloads teachers with a hundreds of curriculum descriptors and there appears little, if any, epistemological justification for the examples given.
>
> (Donnelly 2005, p. 3)

Of course there are criticisms that can be levelled at syllabus models—syllabuses can be badly written, as many were in apartheid South Africa. Even well written syllabuses can be rigid which can cause serious difficulties for children whose development is faster or slower than that predicted by the curriculum. Thus, Green (2000) argues, countries like Japan are looking to the West for models that are more diverse and less prescriptive.

There are clearly different approaches to making judgements about curricula, and much is dependent on the purpose of the evaluation. Some simply rely on the judgement of subject experts. A comparison conducted by the Qualifications and Curriculum Authority in England between A levels and the International Baccalaureate Examinations showed that, while the evaluators expended considerable energy examining the comparability of the two qualifications as a whole, when it came to comparing individual subjects they asked disciplinary experts for their opinions with regards to

- the demands of syllabuses and their assessment instruments (eg question papers, mark schemes); and
- The level of performance required of candidates at key grade boundaries.

(Qualifications and Curriculum Authority 2003).

Experts were drawn from examiners working on the different examinations, as well as from higher education institutions.

Another approach is to provide explicit criteria for evaluating curriculum documents. Donnelly (2005, p. 8) argues, for instance, that the key criterion in evaluating curriculum documents is that they are "clear, succinct, unambiguous, measurable, and based on essential learning as represented by the subject disciplines". When conducting a comparison between Australian and other curriculum documents, he provides the following set of criteria for evaluating curriculum documents:

- identifying key curriculum descriptors, including where first introduced and subsequently dealt with;
- identifying whether the difficulty inherent in the key curriculum descriptors develops across years/levels;
- discussing the depth of coverage of these key curriculum descriptors throughout the documents; including time allocated;
- examining the degree of academic rigour, detail, clarity and ease of measurement of the key curriculum descriptors; and
- noting any significant discrepancies or differences of treatment between the Australian and international curriculum documents.

(Donnelly 2005, p. 4)

A balance between breadth and depth is necessary in any curriculum: a curriculum can attempt to cover too much ground, and as such, fail to deal with any topic in any depth. On the other hand, a curriculum can cover a limited number of topics, each in a lot of depth, and therefore, fail to build a broad base of skills and knowledge. Kerr (2000) says that a good intended curriculum should be focused on what is agreed to be essential (rather than trying to cover everything), specific, manageable for both teachers and students in the time available. It should be focused on conceptual development (rather than long lists of content), sequenced on the basis of evidence (rather than tradition), and supported by shared teacher understanding of what performance at the expected outcome or standard looks like. Finally, he stipulates that a good intended curriculum should be assessable.

As is clear from the very brief foray into the literature, countries around the world are asking what type and model of secondary education should be adopted, and what their curriculum should look like. It is not Umalusi's job to decide on models or even to propose models, but rather to comment on the standard of the existing models, and decide whether or not Umalusi can issue certificates against them. What is clear from both the international literature and our previous research is that it is very difficult to comment on the standard of a curriculum in the absence of externally set and marked summative assessment (Umalusi 2006).

The intended curriculum in South Africa

In South Africa, debates about curriculum prescription are still influenced by the apartheid education system, which was explicitly designed not only to provide separate education for different race groups, but also to indoctrinate children with 'Christian nationalism', and to prepare black children for a role as inferior citizens and as workers (Kallaway 1988). Eighteen separate education departments administered a discriminatory hierarchy of financing, resources, facilities, and quality (Hartshorne 1985). The education provided for black children, referred to as 'Bantu education', has been widely regarded as an attempt to subjugate black people (Buckland 1981; Kallaway 1988). Syllabuses "stressed obedience, communal loyalty, ethnic and national diversity, the acceptance of allocated social roles, piety, and identification with rural culture" (Lodge 1983, p. 116). Syllabuses for the white minority, who had access to better education, were also authoritarian. An often cited-example is the History syllabus, which was seen as "designed to perpetuate an Afrikaner Nationalist interpretation of South African history" (Lowry 1995, p. 106). But while History was particularly open to manipulation, even the Geography syllabus and

textbooks were designed to perpetuate apartheid ideology, giving official recognition to the apartheid landscape and creating particular perspectives about African agriculture as "primitive, irrational, subsistence-oriented and based on low-level technology" (Drummond and Paterson 1991, p. 66).

Syllabuses were developed by the Department of Education with little involvement of any educational stakeholders (NECC 1992). While syllabuses were largely discipline-based, the documents of the intended curriculum were in many instances not clear, rigorous, and concise. There was considerable evidence of syllabus stagnation and neglect, leading to predictability of examinations (Yeld 2005) as well as of syllabuses being in considerable disarray and not well constructed or presented (Umalusi 2006a). Some of these syllabuses were long lists of information with no clear sense of progression or the relationships between content areas.

The ANC-led government which was elected to power in the first democratic elections in 1994 inherited an education system which was "complex and collapsed", with "high levels of adult and matriculation illiteracy, dysfunctional schools and universities, discredited curricula and illegitimate structures of governance" (Chisholm 2003, p. 269). By denying access to education, by providing poor quality education to most black people, by providing poor training to black teachers and by controlling the content of syllabuses to reflect the interests of the apartheid state, the education system reinforced the social and economic inequalities which underpinned apartheid.

The new government, borrowing the idea of outcomes-based education from New Zealand and Australia, introduced an outcomes-based curriculum into the primary and junior secondary school system in 1996 (Spreen 2001; Allais 2007). At the same time, an outcomes-based National Qualifications Framework was developed, which put forward the national prescription of learning outcomes as a mechanism to replace the national prescription of syllabuses (Allais 2007). The version of outcomes-based education initially introduced in South Africa was based on an assumption that the content of the intended curriculum should not be centrally prescribed, but should be developed by teachers against centrally prescribed learning outcomes—in what intended to be a reversal of the Bantu education syllabuses:

> At the heart of Curriculum 2005 is a set of values linked to social justice, human rights, equity and development as well as a learner-centred approach to learning. The intention of outcomes-based education is to improve the quality of the learning experience through methods that emphasise activity-based rather than rote learning.
>
> (Chisholm 2004, no page numbers)

But problems with this curriculum became apparent almost immediately upon its introduction into schools. One of the major criticisms was that the curriculum was excessively complex and that outcomes-based education could only work in well-resourced schools with highly qualified teachers; poorly-qualified teachers in rural schools, it was argued, would be lost when faced with the demands to create their own curricula and resources (Chisholm 2004). Other critics argued that there were aspects of the curriculum which were problematic even for wealthy schools—such as a lack of emphasis on knowledge of basic Science facts and understanding Science concepts in the Science curriculum (Howie 2001).

A review, commissioned by the Minister of Education, suggested substantial

revisions (Curriculum 2005 Review Committee 2000). These recommendations have now been enacted. Because of the problems experienced in the implementation and reviewing of the primary school curriculum, a new curriculum for senior secondary schools was only introduced in 2006. This delay led to a situation which Umalusi has characterized in previous research, using Yeld's (2005) terms, as 'curriculum neglect and stagnation' (Umalusi 2006a). The new curriculum, being phased into secondary schools at the time of writing, contains learning outcomes, but also has specifications of content. Umalusi's recent research (Umalusi 2007) suggests that there are problems with these curriculum documents: some are long and unwieldy for educators to use, particularly because the curriculum documentation consists of three different documents which sometimes contradict one another; also, at times, the documents do not specify the content to be taught sufficiently clearly, and still rely on learning outcomes.

The history of syllabuses and curriculum statements in South Africa remains a difficult and contested one. Clearly, there is still much disagreement amongst South Africans about what kind of intended curriculum we want, how much and what kind of information should be prescribed, how it should be presented to teachers and what teachers should have autonomy over. (See Umalusi (2006c) for a collection of conference papers reflecting some of this contestation.)

Of course there are many factors which contribute to quality in education, and the intended curriculum is just one of them. But clearly, an intended curriculum can be well or poorly constructed, and can do a good or bad job of indicating to teachers what it is that they should be teaching. As part of its role in monitoring standards, Umalusi must ensure that intended curricula are as good as possible.

Umalusi's research into intended and examined curricula 3

Over the last four years Umalusi has conducted three research projects evaluating and comparing different intended and examined curricula. This section provides a brief description of each project.

Project One: Investigation into the Senior Certificate Examination, 2004

The first research project, carried out in 2004, examined the standard of Senior Certificate examinations over a ten year period (Umalusi 2004). For ease of reference, in this document it will be referred to as the *2004 Matric Research.*

Schooling is the largest part of the education system, and the most in the public eye. It was inevitable, therefore, that our first research focus was on the standard of the senior secondary examinations—popularly known as the Matric. The research was designed at a moment when there was public outcry against what was perceived to be a lowering of standards—a perception based partly on a very high pass rate in 2003, and partly on ongoing criticisms from higher education about the calibre of learners entering their institutions. The aim of the *2004 Matric Research* was therefore to compare examinations from 1993, when education was still divided into eighteen different departments; 1999, when pass rates were low; and 2003, when pass rates were at their highest. Subjects examined were English First Language, English Second Language, Biology, History, Mathematics and Physical Science. In other words, we were looking at four different subjects within the *same* qualification, over a period of time, in an attempt to understand whether or not standards were dropping. Evaluators were only asked to make judgements with regard to the standards of the examinations, and not to make evaluative comments about the intended curriculum.

Teams of expert evaluators for each subject were created. Evaluators were supplied with examination papers and marking memoranda for each of the three years in question, as well as marked scripts for *2003*[1]. Evaluators were also supplied with syllabuses, although the aim of this project was not to evaluate syllabuses, but simply to use them as one of the tools with which to make judgements about the examinations.

Umalusi did not prescribe tools for the evaluators but made some suggestions about criteria and categories to guide them. These included content coverage, constructs employed, relative difficulty and challenge presented to learners, variety of task types, length of paper, language and cultural bias, clarity of instructions, organization of paper, additional criteria supplied to candidates, and relationship between paper and marking memorandum. The full set of guidelines is supplied in Appendix Two. Each subject

[1] In some instances slightly different years were selected due to inability to access the required documentation

group developed or adopted a three- or five-point scale to assist them in categorizing the challenge or difficulty level of items.

All evaluators argued that judging levels of difficulty based on question papers and marking memoranda was a highly fraught exercise. Nonetheless, using the suggested guidelines, the scales of levels of difficulty, and their expert judgement, evaluators made clear judgements about the relative standards of examinations in different years. The final published report (Umalusi 2004) presented a synthesis of the findings in relation to each subject, as well as in relation to other aspects of the Senior Certificate examinations system.

Project Two: Apples and Oranges? A comparison of school and college subjects

The second research project compared the syllabuses and examinations of subjects from the Senior Certificate and National Senior Certificate (school and college qualifications respectively). Published as *Apples and Oranges? A comparison of school and college subjects* (Umalusi 2006a), for ease of reference in this document, it will be referred to as the *2005 School/College Comparison*. Here we looked at similar subjects across two different qualifications, examining their syllabuses as well as the 2004 examinations. The qualifications, supposedly at the same level on the National Qualifications Framework, are located in different parts of the education system.

In the *2005 School/College Comparison*, Umalusi required experts to evaluate both examination question papers and syllabuses in their own right. The subjects selected were English (First and Second Language), Mathematics, Science, and Hospitality—four subjects that were part of both the school and college qualifications, albeit in different forms. Two groups of evaluators were created for each subject: a group of practitioners (expert teachers from schools and colleges) and a group of higher education experts. Each group was given syllabuses, examinations, marking memoranda, and a small selection of scripts. The higher education experts were also given the practitioners' reports for consideration.

Evaluators were given a set of guiding questions and categories to assist their work: they were asked to evaluate the intended curriculum, based on their knowledge as subject experts, with regard to key content areas, concepts and procedures. They were also asked to judge the extent to which the different syllabuses prepared learners for higher education. They were given guidelines for evaluating examinations, including sampling and weighting in relation to key content and conceptual areas of the intended curriculum. The detailed guidelines for the evaluation can be found in Appendix Three.

A report was produced (Umalusi 2006a) which argued that while judgements about standards, difficulty levels, and appropriateness of curricula and examinations are difficult and always likely to be imperfect, it is possible to make some judgements about different courses within the same broad subject area. Specifically, the research argued that judgements could be made through utilizing a clear framework which looked at the amount and type of content specified and examined, and the difficulty level of examinations within specified levels of cognitive challenge.

After the completion of this research, Umalusi held a workshop of experts to consider the tools that were utilized and to propose tools that Umalusi could use in

future work. The workshop developed suggestions for sets of criteria and questions as a tool to be used by subject experts when making judgements about curriculum documents. The workshop also made suggestions about *processes* to be used in order to evaluate intended curriculum documents. Suggestions were also made for tools that Umalusi could use when conducting research about standards of examinations, and processes for the development of tools to supplement the current tools for moderating examinations.

The workshop argued that it would be desirable for Umalusi to develop a single tool that could be used across different subjects. Despite considerable debate about the feasibility of this, a proposal was made that the Revised Bloom's Taxonomy (Anderson and Krathwohl 2001) be used as the basis for a single tool that could be utilized across subject areas. It was agreed that Umalusi would test the viability of the approach through research. The full workshop report (Umalusi 2006b) is available on Umalusi's website, www.umalusi.org.za.

Project Three: A comparison of syllabuses and examinations in Ghana, Kenya, South Africa, and Zambia

The proposed tools were then used in a third research project, the 2006 comparison of syllabuses and examinations of four African countries. For ease of reference, this project is referred to in the remainder of this document as the *2006 African Comparison*. In this project we looked at subjects in the same part of the education system (senior secondary school), but across four African countries. The subjects selected were English, Mathematics, Science, and Biology in Ghana, Kenya, South Africa, and Zambia. Syllabuses and the 2004 examinations from each country were examined.

Groups of experts, each consisting of four or five people, including expert practitioners and higher education specialists, all from South Africa, were formed. Evaluators were given the tools which had been developed in the *2005 School/College Comparison* and refined through the workshop discussed above. The tools for evaluating the intended curriculum consisted of a series of essential categories with descriptors. For evaluating examination question papers, evaluators were given similar guidelines to those in the previous research (looking at sampling and weighting), but were also asked to evaluate question papers using the Revised Bloom's Taxonomy (Anderson and Krathwohl 2001; Anderson 2005). The detailed guidelines and tools are provided in Appendix Four.

A report of the project has been compiled, containing findings for each subject as well as overall recommendations for Umalusi, and is available at www.umalusi.org.za.

PROCESSES IN THE THREE PROJECTS

In all three of the research projects, team members initially worked together, and then on their own as individuals. In each project, the initial work included a workshop on the criteria and guidelines for the evaluation. Individual experts produced reports which were then collated by one person in each group into a single subject report. Umalusi, in consultation with the subject experts, produced overall reports based on a synthesis of these subject reports.

A clear progression can be seen in the specificity of the task given to the expert

evaluators in the course of the research projects, with the first project being the most open-ended, and the third having the tightest specifications. A synthesis of these tools and the findings of the three research processes were again presented to a workshop of experts, and further refinements were suggested to both the tools and the processes.

The discussion below reflects on how the evaluators developed and utilized tools in each subject over the course of the three research projects. Aspects of the research processes, and their implications for Umalusi's systems are considered first. The tools used for evaluating the intended curriculum and for evaluating examination question papers are then examined. The discussion also draws on advice and ideas obtained from the two expert workshops.

Reflections on process and systems 4

It was clear from the three research projects that the selection of experts to conduct evaluations, and the ways in which these experts worked together, were key determinants in the quality of the evaluations—arguably more so than the quality of the evaluation criteria. This section therefore reflects briefly on the necessity of the expert-driven process—but also, the problems and future challenges for Umalusi in this regard. The idea of using experts relates to the problem of benchmarking, which is also reflected on. The section then discuses the necessity for Umalusi to carefully consider the types of reports it requires and how it is going to use such reports, as well as the type of documentation that should be requested in order to conduct an evaluation of an intended curriculum. It is also argued that intended and examined curricula need to be evaluated together, or at least in the light of each other, as far as possible.

The selection of experts

In the *2004 Matric Research*, we had groups of three experts for each subject. These groups included a subject specialist, a Grade 12 educator who had produced consistently good results and a university-based discipline or teaching subject expert. In the *2005 School/College Comparison* we had two groups of four evaluators for each subject. The first group was made up of experienced teachers, from both schools and colleges. The second consisted of higher education experts, from traditional universities as well as from universities of technology. An exception was made within the Hospitality group: an expert outside of the area of Hospitality was included because of her expertise in the higher education benchmarking processes. In the *2006 African Comparison*, we aimed to have, in each subject group, three university lecturers, and, either two experienced teachers, or one experienced teacher and one state curriculum expert.

Both practitioners and experts revealed limitations in the judgements that they were able to make. Practitioners were mainly only able to make judgements based on the courses with which they were familiar, and were not always able to see the bigger picture of the subject or discipline as an area of study. However, the experts' knowledge was in itself a limitation from the point of view of judging the user-friendliness of curriculum statements, as they tended to read more into the documentation based on their insight into the subject areas, in ways that the average teacher would not be able to. This is a limitation in an evaluation because curriculum statements are aimed at teachers. Also, higher education experts may have unrealistic ideas about what can be expected of learners in the primary or secondary school system, or in adult basic education or secondary level vocational education programmes.

Thus, it does appear as if both kinds of expertise should be included, and that teachers should specifically be asked to look at user-friendliness of documents. While teachers and higher education experts could be asked to work separately, as was the case in Umalusi's *2005 School/College Comparison*, it would probably be easier and perhaps more fruitful for them to work together, so that they can discuss the curriculum statement from their various standpoints and reach a group decision on it. Groups should contain sufficient numbers to ensure good representation of the discipline or subject area, and it is best for the groups to be made up of an odd number of participants because of the likelihood of differences in judgements. The *2006 African Comparison* used groups of five; this seemed to be a workable number.

It is clear that curriculum statements and examinations can only be judged by experts. The selection of experts is, however, difficult and will continue to be a challenge for Umalusi in the future. For academic subjects, experts are generally located in higher education, but there is a tricky relationship between subjects in further education and higher education. In vocational subjects, experts may be located in industry, but the relationship between an occupation or specific industry and the course which prepares learners for that industry or occupation is even trickier. Who represents the subject? Should Umalusi use education subject experts, or pure disciplinary subject experts? The relationship between the subject and the discipline is not always straightforward—in English and Hospitality in the *2005 School/College Comparison* clear problems emerged.

This study showed that, particularly in a subject like English, the selection of experts is difficult, as the relationship between languages taught at general and further education levels and languages as subjects at higher education is fraught. Should English Literature lecturers be considered as the disciplinary experts? What about linguists? What about applied English lecturers, who tend to focus on critical language awareness? Who is, in fact, the expert on what a primary or secondary English curriculum or examination should look like?

Another challenge is the extent to which experts disagree. Even in Mathematics, there were serious disagreements about what is required and appropriate in a further education level curriculum. Such disagreements highlight the need for panel evaluations to ensure a broad range of perspectives, but panel evaluations will require that Umalusi make difficult decisions at times.

Nonetheless, the use of expertise is vital. What is absolutely clear from this research is that any tools for judging curriculum statements or examinations can never be used in a procedural way, and will at most be guidelines to be used by the subject experts. Any tools that Umalusi develops must strike a balance between capturing critical distinction and being simple enough to be useful. Umalusi should not be under the illusion that developing highly detailed tools will ensure a process that delivers good, accountable results.

It is critical that Umalusi remain aware of the difficult and contested nature of its work. There is a need to be constantly open to expert ideas, and to guard against attempts to introduce rigid procedures into the evaluation processes. Umalusi's legal mandate—to monitor standards—requires it, however, to make decisions, even within this contested area.

What is the benchmark?

In the *2006 African Comparison*, the courses in the study were, in most instances, fundamentally different, and in general, evaluators judged content specification of each in its own terms. Both comparative and individual judgements about content specification raise the issue of benchmarks—should syllabuses be judged against each other?

The notion of a benchmark surfaced repeatedly during the course of the research projects—against what, evaluators asked the Umalusi research team repeatedly, should they be judging curriculum statements? We attempted to partly solve this question through the use of experts—evaluators were told to compare the prescribed curricula against their knowledge of their disciplines. The comparative nature of the research projects was also a partial attempt to deal with this difficulty—seeing what was prescribed in the various curricula highlighted issues for comparison. However, there was contestation in the *2005 School/College Comparison* about Chemistry, with the school evaluators arguing that the additional Physics topics in the college syllabus compensated for the lack of Chemistry, and the higher education evaluators arguing that the absence of Chemistry made the college course a considerably weaker course, despite the additional Physics topics.

Clearly, there is no such thing as 'the ideal curriculum' in any subject. The evaluation of vocational curricula is a particularly good example of this where judgements have to be made about the appropriateness of the selection of knowledge from a particular discipline in relation to a more focused area of study—Mathematics for Motor Mechanics, for example.

In order to evaluate syllabuses, the higher education Mathematics evaluators from the *2005 School/College Comparison* initially proposed to apply the *strands of mathematical proficiency* as outlined by Kilpatrick, Swafford and Findell (2001). However, they later decided to use the outcomes of the new curriculum as an outline of the key areas of mathematical competence that a syllabus must cover. This is worryingly circular, as the curriculum of the new National Senior Certificate is an important one for Umalusi to evaluate—it does not make sense, therefore, for it to be a benchmark.

The Biology evaluators in the *2006 African Comparison* used a selection of textbooks as their benchmark. This raises questions—on what basis is one textbook chosen over another? If textbooks are written based on specified syllabuses, will the exercise not become circular? If international or higher level textbooks are used, will evaluators not make incorrect judgements about legitimate differences in a syllabus based on its aims and context? We should not conclude that it is 'wrong' to use textbooks as some kind of proxy for a benchmark—a textbook can be a useful supplementary tool, provided evaluators select textbooks widely regarded as excellent representations of the discipline, use their professional judgement, and allow for teachers to differ with textbooks where appropriate.

In the *2005 School/College Comparison*, evaluators were asked to comment on whether or not courses prepare learners for higher education. This type of focus could, where appropriate (that is, where courses are aimed at preparation for higher education) provide some type of benchmark. However, even though higher education has lists of 'skills' that it expects from learners, it is not always straightforward to judge whether or not a particular course will impart those skills. This is more of a problem when subjects

at a secondary level are not required in order to study the same subject at higher education. For example, particular mathematical content areas must be mastered by learners in a secondary school level Mathematics course if the learners are going to be able to study Mathematics at higher education. But the same does not apply to History or Hospitality, although both subjects are on the list of 'gateway' subjects which allow access to higher education, and presumably must instil certain generic skills in learners. There is no transparent and non-problematic 'benchmark' for judging whether or not a course will prepare learners for higher education. Some judgements can be made using preparation for higher education as a *guiding principle*—for example, if there is insufficient reading stipulated to prepare a learner for the volume of reading required at a higher education level (as argued by evaluators of English courses in the *2005 School/College Comparison*), or a key content area which is important in higher education is left out (as argued by evaluators of Mathematics courses in the *2005 School/College Comparison*).

Norming against international practice—what other countries expect from learners of a particular age or level of the school system—does, it seems, have some role to play in making judgements, and in this regard, Umalusi should consider obtaining examples of internationally well-regarded syllabuses as supplementary documentation for evaluators.

In summary, there is no fixed benchmark, other than an intelligent and considered look at the discipline in question, keeping in mind the purpose of the course at hand and a reasonable consideration of contextual factors, such as what is appropriate for learners at a particular level of the education system. Where possible, a comparison with other courses said to be at the same or on a similar level is also useful.

Reporting on evaluations

In all three of the research projects discussed above, evaluators' reports differed dramatically, in terms of length, structure, and type of comment. At times, the reports contained aspects that were not necessarily of immediate use to Umalusi—in other words, that did not directly answer the question. Some reports were long, unstructured, and unclear—a lot of work had to be done to process them and work out what they were actually saying. Other reports were succinct and made clear judgements. Umalusi should consider, therefore, how it wants experts to structure their reports and possibly give them more guidance in this regard.

The presentation of curriculum documentation

An issue which emerged from the various research projects is how curriculum documents should be presented. The collection of appropriate documentation from the different countries in the *2006 African Comparison* presented particular problems, as our lack of familiarity with the systems of the three other countries in the study made it difficult to even know which documents to request. While this problem will not be as complicated within South Africa, Umalusi certainly needs to think carefully about the kinds of documents it requires in order to make judgements. This could include giving assessment bodies guidelines about what a curriculum statement should look like. It has already been suggested that Umalusi should obtain either exemplars of examinations or actual examinations as supplementary documentation. Both too little and too much information

will present problems for Umalusi; a balance needs to be struck as clarity must be given to curriculum designers on what documentation is required.

Syllabus evaluations need examinations, examination evaluations need syllabuses

Evaluators emphasized the difficulties of evaluating courses through an analysis of syllabus documentation alone. English evaluators, in particular, emphasized that syllabus documents are always open to interpretation and the ways in which they are understood and implemented vary significantly across school sites and even individual teachers. There are often vast differences between the intended curriculum, as represented by the official syllabus documentation, and the enacted curriculum, as represented by the full spectrum of possible classroom practices. In an examination-based schooling system, these differences can be contained by the examined curriculum since classroom practice is to a large extent delineated by what learners need to know and be able to do in their final examinations. Because this 'backwash effect' of examinations is so powerful—often exerting more influence on classroom practice than the syllabus itself—examinations need to be carefully aligned with the stated syllabus aims or else they run the risk of undermining the very aims which they purport to achieve. The fact that this was emphasized by the English evaluators is perhaps related to the poor state of the old syllabus (teachers currently teach predominantly off the examinations) as well as to the lack of differentiation in the new 'skills-based' curricula, which are particularly difficult to make sense of in the absence of examinations. This interdependence between syllabuses and examinations is clearly the case in all subjects. Science evaluators in the *2005 School/College Comparison* found, for example, that the college (N3) examination covers almost all the areas of the syllabus within 10 possible questions for 100 marks. This leads to superficial coverage of the syllabus and predictability in the exam. Mathematics practitioner evaluators from the 2005 research similarly argued that:

> it is difficult, as an N3 examiner, to set questions that cover all the topics in the tightly specified weighting system and satisfy the stated requirements according to the cognitive levels.
>
> (Kitto, Nkambule, Trollope et al. 2005, p. 9)

Also, there were components in some syllabuses which were not examined at all, and evaluators argued, based on their experience of the system, that they were accordingly not taught. An example is the 'short stories' section in the Business English course. This section is specified in the syllabus, but not examined, and, according to evaluators, is therefore not taught (Allais, Haffejee, Jordt et al. 2005).

Umalusi believes that judgements can and should be made about the quality of the intended curriculum based on syllabuses or curriculum statements as documents in their own right. Nonetheless, it clearly would be useful for evaluators of curriculum statements to have copies of actual or exemplar examination question papers available.

Similarly, it is obvious that syllabuses are required for the evaluation of examination question papers—as question papers have to cover the syllabus. When comparative evaluations need to be made, differences in syllabuses cause differences in standards of examinations—for example, differences in prescribed setworks in language syllabuses affect the relative standards of the corresponding examinations.

As discussed above, the curriculum comparison of British A levels and the

International Baccalaureate conducted by the Qualifications and Curriculum Authority (2003) included an analysis of "the demands of syllabuses and their assessment instruments (such as question papers, mark schemes); and the level of performance required of candidates at key grade boundaries" (Qualifications and Curriculum Authority 2003).

If Umalusi is going to do initial syllabus evaluations in the absence of examinations, the evaluations should be carried out with a fairly 'light touch', that is, they should aim simply to get a sense of whether or not any particular set of curriculum statements in general seem to be adequate for the subject at hand. Umalusi could possibly require assessment bodies to submit a mock examination with their curriculum statement. Umalusi should then supplement its initial report with monitoring of the examinations over time.

Reflections on evaluation of intended curricula 5

This section starts by briefly examining recommendations from the research projects about the tools for evaluating the intended curriculum. It then reflects more broadly on ways in which evaluators made judgements, and what this implies about evaluating intended curricula. Finally, it discusses aspects of some of the categories for evaluation in light of the three research projects.

Recommendations from the research projects

As discussed above, the *2004 Matric Research* did not directly evaluate the intended curriculum. So, the discussion below draws primarily on the *2005 School/College Comparison* and the *2006 African Comparison*. Nonetheless, some evaluative comments about the intended curriculum did emerge in relation to the evaluation of the examinations in the *2004 Matric Research*. For example, the Biology evaluators argued that one of the reasons for the 2003 Biology examination being easier than the 1993 examinations was that the syllabus had been reduced, and there was less content for learners to be examined on (Dempster, Khumalo and Mavovana 2004). English evaluators argued that the removal of the writing examination in English—a move which, on the surface, appeared to be progressive as writing was examined through a portfolio of evidence, in fact meant that less writing was taught. The English group also commented on differences in prescribed set works across provinces, which, they argued, influenced the standards of the different provincial examination papers (Allais, Davidson and Reed 2004; Yeld, Grobler and Sekwane 2004).

Evaluators in the *2005 School/College Comparison* made fruitful analyses of the various curricula based on a comparison of specified content and concepts, as well as assessment specifications. It is not clear that the judgements on syllabuses made by evaluators in the *2006 African Comparison* were a dramatic or substantial improvement on those made by the evaluators in the 2005 research. In both cases, experts were making judgements about the quality of an intended curriculum based primarily on their knowledge of the subject. Nonetheless, the use of more structured criteria and categories for evaluation in the *2006 African Comparison* can be seen to be an advance, both in terms of accountability of expert judgements—as experts made their judgements more explicit in relation to different aspects of the intended curricula, and because aspects of the intended curriculum which were not considered in the *2005 School/College Comparison*, such as coherence of content and presentation and user-friendliness of the intended curriculum were included.

In general, the evaluators in the *2006 African Comparison* felt that the proposed tool

for evaluating curriculum statements provided useful guidelines for Umalusi to use in future evaluations. They emphasized that while a perfect tool is impossible to develop, it is important to have one that is clearly stated at the start of the evaluation. Further, they suggested that the principles underlying such a tool should be as explicit as possible, and emphasized the need for evaluators to spend time together to reach a shared understanding of the various criteria involved.

The Mathematics, Science, and Biology groups did not propose any changes to the tool, which suggests that the tool is fairly successful as a generic set of guidelines. However, the fact that the Biology group introduced two textbooks as benchmarks for their evaluation does raise a question. Their introduction of the books could imply that the tools only work in relation to some kind of external 'benchmark', or indeed that the benchmark, rather than the tool, was primary to the evaluation.

The English group proposed fairly substantial changes to the tool. They argued that these were mainly English specific. However, it seemed that Umalusi could develop a tool based on the recommendations of the English group that could be used for all language courses, and that many of the changes could, in fact, be useful to all subject groups. The proposed changes from the English evaluators were presented at the second workshop of experts discussed above, and further suggestions were made to strengthen the tool, as well as to refine the process for its use.

One critical introduction was a criterion stipulating that evaluators reach a judgement about the user-friendliness of curriculum statements. Curriculum statements are documents which are primarily aimed at teachers, text-book writers, and examiners. If the statements do not clearly and simply inform these groups of people, then the statements do not serve their purpose.

The revised curriculum evaluation tool is attached as Appendix Five, and it is suggested that Umalusi adopt this tool for its evaluations of curriculum statements. Further improvements will undoubtedly be made in the course of its use.

Lessons to be drawn from the research projects

Looking at the two research processes overall, content specification seems to be the most important component with regard to an intended curriculum. The listing of content areas, as well as weighting them in relation to each other, is the central aspect of an intended curriculum that enables evaluators to make judgements on its standard. It was clear from the *2005 School/College Comparison* that where an intended curriculum consists only of broad outcomes, it is not possible to make any evaluative judgements about its standard.

The additional procedural specifications, captured in Umalusi's evaluation criteria, are crucial to ensure that teachers are able to interpret and implement the intended curriculum. For example, teachers should not be forced to decide for themselves how much time should be spent on different sections of the curriculum as this could lead to highly uneven learning across the education system. As such, curriculum coherence, sequencing, and assessment specification emerge as very important factors when examining an intended curriculum.

A very important component of curricula is their aims. They are, however, likely to be, firstly, contested, and secondly, often hard to pin down and rather vague and immeasurable. This is because it is in the aims of a curriculum that the ideology behind

it becomes visible. In this sense, aims are more a political statement than something a curriculum must be narrowly held to. However, the aims specify the purpose of a particular curriculum and could make it clear why a particular selection of content has been made, or how it should be presented. Aims are important, but it is not always easy to make judgements about them.

Cognitive demand proved to be a factor almost impossible to judge in an intended curriculum.

It must be emphasized that it is clear from the research projects, as well as the expert advice obtained in the workshops, that the tool must never be seen as more than a guideline. It is a way of *focusing* the judgements of experts. It should not be seen as a tick-box exercise which evaluators must follow rigidly. Creating the right processes—commissioning the appropriate combination of experts, obtaining considered judgements from a group of experts and so on—will ultimately be more important than developing the perfect tool.

It is worth remembering Donnelly's argument that the key criteria in evaluating curriculum documents is that they are "clear, succinct, unambiguous, measurable, and based on essential learning as represented by the subject disciplines" (Donnelly 2005, p. 8). Perhaps this could be incorporated into a guiding document for curriculum evaluators. Kerr's (2000) argument that a balance between breadth and depth is necessary in any curriculum is also important, because a curriculum can attempt to cover too much ground, and as such fail to deal with any topic in any depth, or can cover too few topics in a lot of depth, and fail to build a broad base of skills and knowledge. Kerr's stipulations that a good intended curriculum should be focused on what is agreed to be essential (rather than trying to cover everything), specific, manageable for both teachers and students in the time available, focused on conceptual development (rather than long lists of content), sequenced on the basis of evidence (rather than tradition), and supported by shared teacher understanding of what performance 'at the expected outcome or standard' looks like, could also be included in a preamble to guidelines for curriculum evaluators.

More detailed consideration of some of the categories

CONTENT AND COMPARISONS

In all the subjects in both the *2005 School/College Comparison* and the *2006 African Comparison*, evaluators created a matrix of topics covered by each syllabus, which assisted them in making quick comparative judgements. As an example, the grid produced by the Science evaluators in the *2005 School/College Comparison* is reproduced below. It is clear that some judgements can be made immediately by looking at this grid—for example, that only the Higher Grade schools syllabus includes all three of Newton's laws and that the college syllabus includes almost no Chemistry. Various other differences in the specification of topics can also be seen. What cannot be judged, however, is the required depth of knowledge, say, in relation to Newton's laws in each of the three intended curricula.

Table 1: Overview of Science content from *2005 School/College Comparison*

Physical Science (HG, schools)	Physical Science (SG, schools)	Engineering Science (N3,colleges)
Physics		
Bodies in Motion: Newton's three Laws of Motion	Bodies in Motion: Newton's First and Second Laws of Motion	Bodies in Motion: Newton's Second Law of Motion
Newton's Law of Universal Gravitation, projectile motion (up and down)	Newton's Law of Universal Gravitation	
Concept of friction	Concept of friction	Friction: static and kinetic friction, horizontal and inclined planes
(Covered in Grade 11, including graphs of motion)	(Covered in Grade 11, including graphs of motion)	Velocity and acceleration, equations of motion, no graphs of motion
(Vectors in general covered in Grade 11)	(Vectors in general covered in Grade 11)	Force as a vector, equilibrium of forces, resultant force, equilibrant, frameworks, roof trusses
		Belt drives and angle of contact
Momentum: as vector, conservation, change in momentum in collisions, force = rate of change of momentum	Momentum: as vector, conservation, change in momentum in collisions	Momentum: conservation
Work, energy and power: concepts and conservation of mechanical energy	Work, energy and power: concepts and conservation of mechanical energy	Work, energy and power: concepts and conservation of mechanical energy
		Moments: turning moment for constant motion, levers and lamina, beams
(Heat: specific heat capacity, transfer of heat covered in Grade 10	(Heat: specific heat capacity, of heat covered in Grade 10	Heat: specific heat capacity, transfer of heat, heat value of a fuel, efficiency, expansion and steam
		Hydraulics: hydraulic presses, work done against a pressure
Electrostatics: electricity at rest, force between charges, electric fields, quantization of charge	Electrostatics: electricity at rest, force between charges, electric fields	
Electric current: current concept, force on current-bearing conductor in magnetic field, force between current-bearing conductors (quantitative), resistance, Ohm's Law, heating effect, power (quantitative), alternating current	Electric current: current concept, force on current-bearing conductor in magnetic field, force between current-bearing conductors (qualitative), resistance, Ohm's Law, heating effect, power (qualitative), alternating current	Cells, simple electrical circuits, electrolysis, Joule's Law, power and energy in DC circuits, alternating current, single-phase transformer
Chemistry		
(Covered in Grade 10)	(Covered in Grade 10)	Elements: constituents of matter, periodic table, metals and non-metals, structure of the atom
Reaction rates and chemical equilibrium, energy of reactions, dynamic equilibrium, equilibrium constant, change of state of equilibrium, equilibrium in solutions, some industrial and other applications	Reaction rates and chemical equilibrium, energy of reactions, dynamic equilibrium, change of state of equilibrium, equilibrium in solutions, some industrial and other applications	

Physical Science (HG, schools)	Physical Science (SG, schools)	Engineering Science (N3,colleges)
Chemistry cont.		
Acids and bases: dissociation of water, pH (quantitative), models for acid and base, acid-base titrations	Acids and bases: pH (qualitative), models for acid and base, acid-base titrations	
Redox reactions: definition in terms of gain or loss of electrons, identifying oxidising and reducing agents	Redox reactions: definition in terms of gain or loss of electrons, identifying oxidising and reducing agents	Redox reactions (brief introduction) and corrosion
Electrochemical cells: copper-zinc cell, electrolysis and electroplating	Electrochemical cells: copper-zinc cell, electrolysis and electroplating	Electron transfer: formation of ions, brief definition of electrolysis and electroplating
Half-cell potentials: table of redox half-reactions and applications, selection of reference electrode, calculations of potential difference	Half-cell potentials: table of redox half-reactions, use of table to balance redox half-reactions	
Organic chemistry: definition, structure, nomenclature, hydrocarbons, alkyl-halides, alcohols, carboxylic acids	Organic chemistry: definition, structure, nomenclature, hydrocarbons, alkyl-halides	

Source: (Grussendorff, Makhafola, Reeves et al. 2005)

On the basis of this grid, evaluators were able to show that

> the matric HG syllabus goes into the most depth in all content topics. The SG course covers a similar set of topics, although somewhat reduced in depth and breadth. In the N3 course there are clear and serious omissions, although this syllabus covers a greater number of specific industrial applications.
>
> (Grussendorff, Makhafola et al. 2005, p. 4)

All subject evaluators were able to make some evaluative judgements on the basis of this type of grid of topics. English evaluators, for example, were able, at a glance, to point out significant differences between courses—for example, college courses lacked literature and extended reading, as well as creative and academic writing. The grid is thus a useful tool from a comparative point of view and may also be useful when evaluating individual curriculum statements, as it provides a quick overview of topics in the syllabus.

Weighting of content areas was something that several of the evaluators commented on. A link to the weighting of each topic may be helpful so that evaluators can understand the relative importance of each content area. This also implies that syllabus documents or curriculum statements should indicate very clearly the weighting of different areas of the curriculum statement. Hospitality evaluators from the *2005 School/College Comparison* pointed out that

> unsatisfactory weighting of different modules has led to confusion as to know how much time had to be spent on these modules. Since official directives from the department are minimal, most of the activities regarding this subject are transferred orally from school to school.
>
> (Lotz, Steyn, Venter et al. 2005, p. 3)

Evaluators not only made comparative judgements, but also discussed what they thought should be in a syllabus at this level. For example, the Mathematics evaluators argued that

> Sequences and Series are a worrying omission from the N3 syllabus in that they are a fundamental topic in mathematics and are important in calculus at higher levels. They are a specific kind of pattern, and function and flexibility in working with them adds to students'

> mathematical knowledge and skills in these areas. It is also surprising that problems on maxima and minima are left out of N3, given its more practical orientation and the importance of similar problems at higher levels.
>
> (Brodie 2005, p. 3)

CONTENT IN ENGLISH

While an evaluation of prescribed content was key in terms of evaluating intended curricula, content in English repeatedly surfaced as a problem during the course of the three investigations. In the *2004 Matric Research*, both English First Language and English Second Language evaluators felt that that "it was difficult to establish what content should be covered and whether such content had in fact been reflected in an examination question paper" (Umalusi 2004, p. 32).

> The practitioner evaluators for *2005 School/College Comparison* also argued that English First Language treats English as both a subject with its own content and also as a means of communication. NQF and Business English seem to regard English only as a means of communication. English as a communication tool is likely to be less cognitively challenging than English as a subject as can be seen in the data.
>
> (Allais, Haffejee et al. 2005, p. 10)

The Higher Education evaluators from the *2005 School/College Comparison* contested the idea of language as a means of communication, but did not agree with the analysis of English content that the practitioners described. They instead recommended what they called a 'practices' approach, arguing that a higher education discipline's literacy practices, a corporate business's literacy practices, or a hairdressing salon's literacy practices are not context-free sets of language skills, but rather a set of discourses determined by the norms and values of the community of practice:

> … when literacy is conceptualised as literacies located in a range of social practices, each of which is embedded in a specific context, then it is no longer possible to separate literacy from the people who use it. Researching literacies therefore involves seeking an understanding of the groups and institutions that socialise people into specific literacy practices.
>
> (Jaffer, McKenna and Reed 2005, p. 4)

They argued that by implication a good English syllabus is one which both problematizes the literacies that students are being inducted into, and which also makes them more overt, for easier acquisition.

The evaluators for the *2006 African Comparison*, on the other hand, while criticizing the idea of language curricula in which language is seen as objective and politically neutral, still advocated a 'skills' approach to language teaching. They argued that English is inherently skills-based. However, they criticized the outcomes-based curriculum statements of the new National Senior Certificate in South Africa, arguing that there is hardly any differentiation between different language levels and between different grade levels. This is an inevitable consequence of a curriculum statement or syllabus which focuses exclusively on 'skills' or 'outcomes' at the expense of content. This brings us back to the question of what content is required in an English language course. All the evaluators agreed that learners needed to be exposed to reading of continuous prose in a range of different genres. Perhaps, therefore, a useful distinction would be between a

text-based and a skills-based curriculum. Within a framework of text-based curricula for languages, Umalusi's evaluators could look at the types and number of texts in the intended curriculum. This can apply to the texts that learners are supposed both to read and to produce, and it does not remove the possibility of a critical focus on how language is used in the texts, or a focus on explicit language practices such as grammatical rules or linguistic conventions. In the absence of such a content framework, however, it is not clear how Umalusi can make judgements about the intended curricula of English courses. It is essential that judgements be made, however, as language is clearly very crucial for all our learners.

Evaluators in the various groups made arguments about the relative merit of including or excluding various content areas and did have irreconcilable differences on some matters. While this type of difference is inevitable, Umalusi will have to find ways of dealing with it.

COHERENCE, ORGANIZING PRINCIPLES, SEQUENCE AND PROGRESSION

Coherence, organising principles, sequence, and progression emerged as important factors in the evaluation of the intended curriculum. The English evaluators from the *2006 African Comparison* felt that these factors should be merged, arguing that the notion of 'organizing principle' implies the structure and direction alluded to by 'sequence', and that 'pacing' is simply the speed at which the sequenced syllabus is covered and how the year-end points are positioned. They also questioned the idea of 'organizing principle', arguing that it was difficult to apply it to English. However, experts in the workshop which considered the research findings argued that it is essential that a syllabus or curriculum statement has a clear organizing principle, and that it is coherent. It was also agreed that sequence and progression should be evaluated separately as important aspects of an intended curriculum.

> Evaluators did make evaluative judgements about the overall coherence of the syllabuses. For example, the Mathematics evaluators from the *2005 School/College Comparison* argued that both the school and college syllabuses are inadequately structured.

Specification of content over years (an aspect of sequencing) was also an issue. For example, the lead English practitioner evaluator from the *2005 School/College Comparison*, commenting on one of the new NQF-aligned courses, argues that

> The NQF requires 'bias', and 'semantics' for Level 3, not for Level 4. As an experienced educator, I would make the comment if concepts like this are to be internalised by students, they need frequent exposure and revision.
>
> (Allais, Haffejee et al. 2005, p. 9)

Evaluative comments were made on sequencing, with evaluators suggesting that sequencing should be related to structuring and organizing principles. It must be noted that sequencing is very differently viewed and applied in different areas. For example, it is striking from the detailed discussion of the five Biology syllabuses in the *2006 African Comparison* how different the sequencing is across the countries in question. It almost appears as if there is no necessary sequencing in Biology. As such, it can be seen that, while sequencing and progression are important for an intended curriculum, Umalusi must allow subject experts to have flexibility in terms of how they make judgements about these aspects.

AIMS

As discussed above, the notion of curriculum aims is a difficult and contested one. Evaluators in both the *2005 School/College Comparison*, and the *2006 African Comparison* felt that a consideration of aims provided an understanding of important differences between courses. For example,

> the first general aim of Business English [the college English course] is that 'Students must be able to communicate ideas, thoughts and feelings effectively through written and spoken language in such a way that they can take their place in the community and *function productively in their working environment*.' … By contrast, the first general aim of the English First Language syllabus 'is to help pupils to develop their potential as individuals and as members of society through developing their competence and performance in using language and through enriching their experience and enjoyment of language.'
>
> (Allais, Haffejee et al. 2005, p. 10)

It is important, all evaluators argued, that a syllabus must be judged on its own terms and based on what its aims were. Science practitioner evaluators in the *2005 School/College Comparison*, for example, argued that the school and college Science courses were 'different but equal'—although they had different content specified, each was appropriate in terms of the general aims of the course. However, the higher education evaluators rejected this analysis, arguing instead that the school course was of a higher standard than the college course, because of the amount, type, and level of the specified content, as well as the assessment specifications. The English evaluators commented that because the important content and skill areas of English in the Business English syllabus were dealt with in relation to the aim of preparing learners for the workplace, the syllabus was overly instrumental and narrow, which was, ironically, counter-productive to its intentions:

> If workers are to be life-long learners, and update themselves continually in theoretical aspects of their jobs, they will need reading, and reading skills are best learnt in an atmosphere of reading for pleasure and relaxation. This aspect is completely missing from both Business English and NQF. Whereas arguably the English First Language syllabuses do not require sufficient reading, neither NQF nor Business English courses even require a book to be read.
>
> (Allais, Haffejee et al. 2005, p. 10)

> In terms of developing entrepreneurial skills (a critical outcome), whereas all syllabuses may be wanting, there is no evidence that anything in Business English or NQF teaches learners to think laterally, creatively or conceptually. Some literature, some of the more challenging questions in First Language, some of the creative writing required (although not seen in this study which was limited to examinations) encourage creativity and lateral thinking.
>
> (Allais, Haffejee et al. 2005, p. 11)

In addition, the notion of simply evaluating a course according to its own specified aims is problematic, because its aims may be limited, or inappropriate. Another difficulty with aims is that, as pointed out by the Mathematics evaluators from the *2006 African Comparison*, admirable aims can be stated in a syllabus, but the syllabus can still be designed in such a way that it is not clear how the aims can be achieved. This is particularly the case with aims such as 'preparing the learner for active citizenship.'

Evaluators can, however, look holistically at intended curriculum documentation, and evaluate not only whether or not the stated aims seem laudable in their own right, but the extent to which a curriculum statement is likely to achieve those aims.

COGNITIVE COVERAGE BY COGNITIVE DEMAND

Some of the evaluators attempted to make judgements about difficulty levels within intended curricula. For example, evaluators for English First and Second Language in the *2005 School/College Comparison* attempted to evaluate key areas of the syllabus in terms of difficulty levels. The English First Language evaluators classified Literature as a generally difficult area of the curriculum. This classification was corroborated by the analysis of the examinations, in which learners seemed to find the literature sections to be the most challenging. Generally, however, this was not perceived to be a fruitful exercise. Most areas of the curriculum were classified as potentially easy or difficult, however, once evaluators had looked at examinations, they found that syllabus areas which looked similar were in fact examined at very different levels across the two qualifications in the study (Allais, Haffejee et al. 2005).

The evaluators' judgements about the content and skill coverage of the syllabus, as well as their general analyses of the syllabuses, were more useful. Evaluators were able to make clear judgements on the adequacy of syllabuses, based on, for example, the number and type of prescribed texts, and the number and type of writing assignments.

Evaluators felt that it was impossible to judge levels of cognitive demand on the basis of syllabus documentation. For example, one of the syllabuses examined contained a section on calculus, an area of Mathematics which, they argued, contains difficult concepts. However, a scrutiny of the examination showed that, in fact, only the most cursory and introductory aspects of calculus were tested.

ASSESSMENT SPECIFICATIONS

Looking at the assessment format—including aspects such as the number of papers, type of papers, length of papers, rigidity of mark allocation per content area, rigidity of format and so on—enabled examiners to make more informed judgements about the intended curriculum. For example, evaluators of both the Mathematics and Physical Science in the *2005 School/College Comparison* argued that the assessment specifications for the college courses, including the fact that there was only a single examination which was written over a short period of time, meant that the intended curriculum was likely to be taught in a very shallow way.

Clear assessment specifications need to be a prerequisite in curriculum statements.

TIME

Although it was not specified as a category in any of the research projects, evaluators considered notional learning hours as well as length of contact time between educators and learners when making their judgements about syllabuses in both the *2005 School/College Comparison* and the *2006 African Comparison*. For example, higher education Science evaluators from the *2005 School/College Comparison* argued that

> There is clearly a vast difference between the Matric Physical Science courses and the N3 Engineering Science course in terms of time spent on tuition and concept development.
>
> (Grussendorff, Makhafola et al. 2005, p. 2)

This difference is shown in Table 2 below, taken from the Science practitioners' report (Chapman, Karsan, Mohlamonyane et al. 2005, p. 2).

Table 2: Summary of syllabus coverage in planned time

	SC (Grade 12) Physical Science Paper 1	N3 Engineering Science
Total duration	1 year	13 weeks
Total teaching hours	104 hours	60 hours

Even this comparison is not strictly accurate, because in fact grade 11 content is examinable in grade 12, making the school course a two-year programme.

The Science evaluators from the *2006 African Comparison* argued that the separate Physics and Chemistry courses in Kenya, Ghana, and Zambia contained the same amount of content as the single combined Physical Science course in the new South African curriculum statement. This, they argued, made it likely that either the course would be unrealistically demanding or that topics would be dealt with very superficially.

The notion of weighting was discussed above in relation to content. While weighting or credit value are useful concepts, it is suggested that the time prescribed for mastery of the intended curriculum could be examined as an important aspect of that curriculum.

While some experts suggested to us that the notion of credit is more useful than time, especially because distance education courses could be mastered in different amounts of time depending on learners' abilities, all evaluators were critical of courses which did not stipulate time to be spent on different aspects of the curriculum. Particularly because curriculum statements or syllabuses are supposed to be a useful guideline for teachers and, therefore, time specifications are required.

THE 'IDEAL SYLLABUS' VERSUS CONTEXT

Some of the evaluators argued that while tools or guidelines are required to make judgements, this does not imply that there is an abstract set of principles that constitute an ideal syllabus or curriculum statement. These evaluators felt that the strength of any curriculum statements relates to their suitability for the context for which they were designed. The English evaluators from the *2006 African Comparison* in particular argued that one of the principles of a good syllabus is the extent to which it responds to the needs of its constituent student population and those of their teachers, and suggested that this principle could perhaps be included in the guidelines for evaluating curriculum statements. This point is contested, and could lead Umalusi into the dangerous waters of relativism. But if the other points above are born in mind (particularly, that any intended curriculum must be based on essential learning as represented by the subject or discipline in question) then the suitability of a curriculum for the student and teacher populations, particularly how user-friendly the documents are for teachers, can be seen as an important additional consideration.

HOW THE CURRICULUM WAS DESIGNED

This criterion needs to be added to ensure that appropriate consultation has taken place in the creation of curriculum statements, and to ensure that processes for the creation of curriculum statements are systematic. This is particularly important in South Africa, with both the apartheid history of very authoritarian syllabuses and the consultative approach taken with the outcomes-based curriculum statements. The designers of the curriculum should be identified in the document.

Summary of recommendations on evaluating the intended curriculum

Umalusi should use the tools in Appendix Five as guidelines for the evaluation of curriculum statements. However, the tools must be seen as guidelines and not as a rigid list of items to be ticked off by evaluators.

Evaluators must be chosen based on their expertise in the respective discipline or subject, as well as their knowledge of the part of the education system for which the curriculum statement is intended.

Panels of evaluators should be composed of evaluators with complementary forms of expertise, and at least part of the evaluation process should take place with all the evaluators present.

Evaluators should consider the aims of an intended curriculum and should consider the extent to which the curriculum documentation makes it likely that the aims will be achieved.

Where possible, curriculum evaluators should be given examples of internationally well-regarded syllabuses as supplementary documents, although they must be instructed not to make narrow comparisons against them. Evaluators could also use other supplementary documentation, such as credible textbooks. However, it must be accepted that there is no benchmark, other than an intelligent and considered application of the discipline, the purpose of the course at hand, and a reasonable consideration of contextual factors, such as what is appropriate for learners at a particular level of the education system.

Key content areas and information about how they are weighted must be clearly specified in curriculum statements. Evaluators must be able to make judgements about curriculum statements in this regard. Umalusi needs to find ways of dealing with serious disagreements amongst evaluators.

Umalusi must develop a document for curriculum developers, based on the tool in Appendix Five, to make it clear to them what is required of their curriculum statements or syllabuses.

Curriculum statements should contain guidelines for how much time should be spent covering the various aspects of the intended curriculum.

Umalusi should request that assessment bodies supply either mock examinations (for new curriculum statements) or examples of actual examinations (for already existing courses). These should be used as supplementary documentation for evaluators. In the case of new curriculum statements, Umalusi should monitor the quality of examinations over time, and revise its judgements about the intended curriculum if necessary.

Umalusi must specify clearly to curriculum designers what documentation should be submitted for evaluation, bearing in mind that the point of evaluations is to ensure that clear and succinct documentation is provided to teachers, textbook writers, and examiners.

Evaluating examination question papers 6

As discussed above, in each of the three research projects, evaluators were provided with guidelines and categories for evaluation. In each project, evaluators also developed grids for evaluating levels of cognitive challenge. The broad guidelines were similar to those used by Umalusi's moderators, and, in general, were seen to be useful. The discussion below first reflects broadly on what can be learnt from the three projects, and then focuses more specifically on the grids used by the different subject experts in attempting to make judgements about levels and types of cognitive challenge in the different examinations. Two other factors contributing to the standard of examination questions papers—predictability and marking, are then considered.

Thinking about cognitive challenge across the projects

It is clear from all three projects that evaluators felt the need for some kind of hierarchy to evaluate individual items. All evaluators worked with the idea of levels of difficulty, although, these were differently described, in increasingly complex ways, across the three projects. Grids of hierarchy where developed. Two things became clear from the research: firstly, that a grid with two axes is difficult to work with, and secondly, that individual subjects require their own grids.

The groups in the first two research projects developed tools when they were not given any. The tools consisted of simple grids with easy, medium, and hard designations. There were some descriptions of what makes a question and required task easy, medium, or hard, and some examples. The *2004 Matric Research* also entertained the idea of using the Revised Bloom's Taxonomy, because it identifies skills and abilities in the knowledge processing hierarchy, from the highest to the lowest level of complexity and difficulty. The 'structure of the observed learning outcomes' (SOLO) taxonomy was also examined as a possible set of tools. This taxonomy categorizes learners' understanding in terms of structures of knowledge—pre-structural, uni-structural and multi-structural, relational as well as structural knowledge (Biggs and Collis 1982). However, this type of approach seems more geared to ongoing evaluation of learners than once-off exit examinations, and was not taken up by any of the groups. Bloom's Taxonomy has been used in Umalusi's moderation criteria (and is used in the examination specifications of the other African countries in the 2006 study).

Based on an analysis of the two projects, Umalusi decided to attempt to distinguish not only between levels of difficulty but also types of cognitive challenge, and wanted to see if it was possible to have a generic tool across all subjects. Thus, the third research group were given the Revised Bloom's Taxonomy, in an attempt to reach more

nuanced judgements, and in an attempt to make it possible to compare difficulty levels of different subjects. But except for the Biology group, evaluators changed the provided grid, dramatically simplifying it, and making it more appropriate for their subject in various ways.

In all the projects, evaluators weighed up mark allocation against the various categories that they were using for judging levels of cognitive challenge.

The evaluators made holistic judgements based on balancing a range of different things: syllabus coverage, weighting, type of question, type and length of answer required, predictability, challenge level, overall length of paper, clarity of presentation, and amount and type of reading required. For example, the Biology evaluators from the *2004 Matric Research* pointed out that:

> The many short questions permit a fairly superficial coverage of all topics, *but do not probe learners' ability to articulate their answers in a written form, nor to illustrate their answers appropriately*. A further disadvantage of breaking the questions down into many small questions is that *opportunities to test ability to synthesize information from different topics are limite*d. General principles underlying the biological processes occurring in plants and humans are rarely explored. An example of such a question would be "Compare and contrast the gaseous exchange surfaces of plants and humans."
>
> (Dempster, Khumalo et al. 2004, p. 14, emphasis in original)

This type of analysis is more textured and probably more useful than simple categorization of questions into levels on a grid.

Looking all the projects, it is clear that it is hard to reach clarity about *challenge and difficulty level* and, as such, criteria were elusive. There was considerable debate in the subject teams about the meaning of challenge or difficulty levels in their respective subjects. For example, the Mathematics practitioner evaluators from the *2005 School/ College Comparison* argued that:

> It should be mentioned that candidates do not always do best on questions classified as level 1. A case in point is that candidates are often unable to reproduce the proofs of theorems in the Euclidean Geometry section of the Senior Certificate curriculum (which is not included in the N3 curriculum). Some teachers put this down to lack of application by the pupils, whilst others believe that the cognitive demands of learning proofs which contain some fairly complex concepts and logical arguments, are beyond many/most Standard Grade candidates.
>
> (Kitto, Nkambule et al. 2005)

The result of this debate in each group was a series of definitions coined by the groups, or adapted from curriculum documents, to explain what they meant by challenge or difficulty levels in their respective subjects. There was no simple consensus on what constitutes challenge or difficulty levels. Evaluators argued that it is very difficult to decide precisely what skill, competence, or ability is elicited by a particular item. Nonetheless, within groups there was a high degree of unanimity among evaluators in their judgements about the overall standards of question papers. This could imply that their judgements were primarily based on intuition and experience in the subject and how it is taught and tested rather than on tools or criteria.

What do these three research processes tell us about making judgements about the level of cognitive challenge of examination question papers? It seems that there are various possibilities. It could be the case that the Revised Bloom's Taxonomy is not the

right generic tool, but an appropriate generic tool can be developed or found. It could be the case that no generic tool will be useful. It could be the case that any tool, whether generic or subject specific, will be no more than a rough guide, and will be imperfect.

It does seem, however, that having a tool is useful. In each of the research processes, a hierarchy of levels of cognitive challenge seems to have helped to focus the minds of the evaluators. All the groups seemed to have needed a sense of a hierarchy. But putting an exact label or description on that hierarchy may prove to be an unreachable target.

The tools got more and more explicit through the course of the three projects, but it is not clear that the quality of judgements was dramatically improved, and it is possible that a more detailed tool could distract experts from their own sense of judgement. For example, taxonomies which focus on different types of verbs could lead evaluators to simplistically categorize questions which begin with those verbs, even though a question which says 'explain' could in fact be a recall question in disguise.

It is worth noting here that evaluators in all three projects found significant variations across provinces—something that vindicates the Department of Education's decision to set more papers nationally. This was aggravated in English by substantial differences in prescribed texts, particularly in English Second Language, and it is not clear whether or not this will in fact be remedied in the new curriculum. The higher education evaluators from *Apples and Oranges*? pointed out that while the issue of equivalence is a difficult one across the education system, it is particularly problematic that there are substantial differences within examinations based on the same syllabus. Evaluators also repeatedly found that examinations from the IEB were of a higher standard than those of the state. The differences across examination bodies suggest that it is very difficult for moderators to achieve standardization up front in the evaluation of question papers.

All of this reinforces the point that making judgements about the difficulty level of examinations beforehand will never be exact. Umalusi may be better off conducting analyses of learners' performance on items after writing, than in putting too much effort into the elusive search for the perfect moderation tool. Nonetheless, it is important that moderation of question papers is done as well as possible. As such, it is recommended that Umalusi should consider including the subject specific grids used in the research projects as additional tools for moderators. In English, the grids used in the first two projects seem to have been more productive than those used in the third project. In Science the grids from the higher education evaluators from the *2005 School/College Comparison* and from the *2006 African Comparison* (Tables 10 and 11 in the Science section below) both seem to have been useful, and are more nuanced than a simple three level grid. In the other subjects, one of the various grids produced could be adopted. Umalusi may need discussion with curriculum designers in this regard, as many curriculum statements contain grids specifying levels of performance expected from learners, or rubrics for evaluating learner performance. It would be desirable for such tools to be in harmony with tools used by Umalusi wherever possible.

What emerged very clearly from this research is that the tools were not sufficiently transparent, even with examples, although evaluators found examples very helpful. All the evaluators pointed to the need to spend a considerable time working together in order to build joint understanding. This highlights the need for panel evaluation; with the

panel working together at least part of the time, as well as the building up of expertise over time.

The discussion below presents in more detail the specific tools used by evaluators in each of the three research processes, and discusses the ways in which evaluators were able to use them.

ENGLISH

English as a subject was common across all three projects. In the first project, Matric English First and Second Language were evaluated across the ten-year period. In the second project, Matric English First and Second Language were compared with Business English, the English course offered in the colleges. In the third project, Matric English First and Second Language were compared with various English courses in the three other African countries in the study (while all three countries have a compulsory English course, Ghana and Zambia also have elective English Literature courses). Within South Africa, evaluators looked at the old syllabuses (English First and Second Language) as well as the accompanying examinations and also examined the new curriculum statements (English Home and English First Additional Language).

In the *2004 Matric Research*, the English Second Language group created a three-level grid differentiating between simple, relatively simple or challenging questions (Yeld, Grobler et al. 2004). They also considered other variables, such as the competence or skill covered, and the nature of the task or way the competence or skill was elicited (the item type). They argued that:

> [i]n Language, what determines the standard of an examination and its degree of difficulty is less a matter of whether a syllabus has been 'covered' in the examination, and more to do with two aspects of the paper. Firstly, the verbal and visual texts selected for comprehension, summary and language work, and secondly, the degree of challenge of the questions which are asked.
>
> (Allais, Davidson et al. 2004, p. 1)

The table below is the resulting grid, which they supplied with examples.

Table 3: Grid used by English Second Language evaluators in 2004 Matric Research

	Options	Examples
1	The task was very simple. The context/text of the task was very simple. The task was relatively complex but was routinised/predictable.	How many paintings does Portchie produce every year?
2	The task was relatively complex but the context/text made it less so. The task was simple but the context/text was fairly elaborated or complex.	A quite challenging 'word search' item where the word would have been unfamiliar but the context made it obvious (the desired word stuck out like a sore thumb), or the distractors in a MCQ item were obviously incorrect. An active into passive item with a number of transformations.
3	The task itself needed some unpacking – it was not always clear what was wanted – would require some strategic thinking. Complex task involving more than one operation. Task to be performed on a complex or subtle text.	

(Yeld, Grobler et al. 2004)

They used this grid to categorize all the items in the papers in the study, factoring in the mark allocation. So, for example, an item worth 3 marks which had been assessed as at challenge level 2, scored a 6 (3 x 2). They argued that locating the challenge level in the task rather than in the *skill* helped the group to achieve consensus, even though they felt that this was a very crude proxy for task analysis and cognitive functioning.

The grid developed by the Second Language evaluators in the *2005 School/College Comparison* was similar, although in some ways less developed:

Table 4: Grid used by English Second Language evaluators in *2005 School/College Comparison*

Level scale	Descriptor/s of cognitive requirements
1	**Simple task/context, knowledge, recall, literal comprehension, low demand, lower order thinking skills e.g.** "List three ways in which children are protected by the anti-smoking legislation." (Grade 12 , HG, Paper 1, question 1.3)
2	**Meaning translation, inference, context/task more advanced, analysis and application of knowledge, average skills e.g.** "In paragraph nine there is a phrase 'to live "in sin" for 20 years'. In the context of this article, what does this phrase mean?" (NSC, Paper 1, question 1(18)).
3	**Reorganization, synthesis, evaluation, appreciation, context/task complex, requires strategic and discriminative thinking skills e.g.** "Rewrite the following sentences.The doctor told her to lean back and look at him. She recognized the doctor. Begin as indicated: Only after..." (Grade 12 HG, Paper 1, question 5.4).

(Grobler, Akaloo, Sekwane et al. 2005)

Again, in deriving the challenge level, the score was factored in. An item worth 3 marks which was assessed at challenge level 3 scored a 9 (3 marks awarded for question x 3 for difficulty level). Each question within a section, for example, comprehension or language, was rated and a difficulty factor calculated for the section as a whole. This process was not considered an end in itself but rather as a basis for comparison between parts of papers where similar skills were tested.

The English Second Language evaluators for the *2004 Matric Research* also produced a more detailed set of tools for making judgements specifically about summary questions, as shown overleaf.

Table 5: Framework for the analysis of the summary

Assessment of text/s used in summary
Genre
General accessibility and interest of text
Audience sensitivity (gender, 'race', culture etc.)
Length (250-300 words)
Vocabulary level
Sequence of events (e.g. chronological order)
Fitness for purpose (i.e. is the text suited to the assessment of 'real-life' Language-Across-the-Curriculum summarizing needs of students?)
Task requirements
Selection, condensation, production – does the candidate need to make choices about the relative importance and relevance of pieces of information, etc.? – How difficult is it to identify points? – How difficult is it to rephrase points to suit the leading question?
Super-ordination – does the candidate need to develop a conceptual framework to organize the information needed to complete the task?
Transformation – in producing the summary, does the candidate have to construct significantly new text or can s/he simply paraphrase selected pieces of information?
Clarity of instructions
Are instructions clear?
Overall assessment of summary task

(Yeld, Grobler et al. 2004)

This additional framework was developed in order to facilitate analysis in a way that would lend itself to comparison. Evaluators also felt it assisted them to probe the extent to which the summary tasks and texts contained in the various examination papers provided opportunities for the processes of super-ordination and transformation to be used and assessed, along with the more standard summary task characteristics. These tasks, they argued, are key academic skills.

The English First Language evaluators from the *2004 Matric Research* used a three-point scale to describe the level of difficulty:

> We decided to rate individual questions according to cognitive and subject-specific challenge on a scale of (1)-(3), with (1) indicating the simplest and most basic question types and/or the knowledge required to answer them, (2) indicating questions of average difficulty for school-leaving English first-language speakers, and (3) involving fairly sophisticated linguistic, literary and general knowledge.
>
> (Allais, Davidson et al. 2004)

This is very similar to the three-point scale developed by the English First Language evaluators in the *2005 School/College Comparison*:

> **Level 1** requires straight recall, knowledge or literal comprehension.
>
> **Level 2** requires a little more thinking, application or analysis. Candidates might be required

> to give a reason for their answer, paraphrase a passage or even only comprehend literally but in a more difficult context such as literal comprehension from Shakespeare as opposed to literal comprehension from a simple current magazine article.
>
> **Level 3** requires more conceptual thinking. Synthesis, evaluation, and making inferences might be required.
>
> (Allais, Haffejee et al. 2005)

Although evaluators argued that allocating difficulty levels to items was not straightforward, and in some instances very contested, they were able to make judgements about items as well as about question papers over-all, and reach fairly clear decisions about the items. An example is shown below in Table 6.

Table 6: Summary of percentage of difficulty levels in English language papers

	School, higher grade paper	School, standard grade paper	College paper
Paper 1 Language	13% level 1 158% level 2 32% level 3	29% level 1 51% level 2 15% level 3	88% level 1 12% level 2 0% Level 3

(Allais, Haffejee et al. 2005, p. 5)

Evaluators were thus able to argue that despite syllabuses which did not differ radically (with the exception of literature and some types of writing), the college qualification in fact expects a lower standard from its learners.

Clearly, both groups developed grids which they found workable. It is conceivable that the process of developing the tools as well as applying them in a group improved evaluators' ability to make rigorous judgements. The use of grids did not mean all points of contention were eliminated.

In the *2006 African Comparison*, the English evaluators decided not to utilize the *Revised Bloom's Taxonomy* proposed by Umalusi. They provided strong arguments as to why it is not applicable to English, and why they found it to be unworkable (Umalusi 2007). Instead they returned to Bloom's original taxonomy, made some minor adjustments, and added a vertical dimension that specified skills/content coverage, as seen below in Table 7.

Table 7: Grid used by English evaluators in *2006 African Comparison*

		Straight recall, literal comprehension	Developing an interpretation	Applying existing knowledge to a new context	Analyzing	Synthesizing Creating Combining	Evaluating, making judgements	Total marks
Reading	Making meaning of non-literary texts in continuous prose							
	Making meaning of non-literary visual and verbal texts in non-continuous prose							
	Making meaning of literary texts of various genres							
Language structures	Working with grammar and other linguistic features							
Writing	Extended writing according to particular genre or form demands							

(Ferreira, Davis, Distiller et al. 2006)

The resulting grid/matrix was useful in providing a general impression of cognitive demand across the horizontal axis against key content and skill areas down the vertical axis.

However, the evaluators were not convinced that there exist clear, uncontested distinctions between the various cognitive processes specified by Bloom, particularly in English, and argued that this grid did not provide quantifiable, statistical 'evidence' of an examination paper's quality (Ferreira, Davis et al. 2006). Umalusi had requested that evaluators allocate difficulty levels to items, but the English evaluators were against this, arguing that there is a general increase in level of cognitive demand moving across the grid from left to right. In their discussion of the papers, however, they referred to an examination paper as assessing predominantly lower (recall/literal comprehension/basic interpretation), middle (interpretation/application/analysis) or higher order (analysis/synthesis/evaluation) cognitive processes rather than providing a detailed breakdown of mark allocation pertaining to each individual cognitive process.

In other words, despite their attempts to use this ostensibly more sophisticated grid, in practice evaluators fell back on something similar to the three-level grids of evaluators in the previous projects.

Evaluators also argued that the mapping of an examination paper onto a grid needs to be supplemented by close consideration of broader questions pertaining to sampling and weighting. In this regard, they felt that the guiding questions provided by Umalusi were useful. They suggested minor changes to these questions, which could be incorporated into moderation tools. They also highlighted three additional issues:

- with regards to sampling, text selection needs to be considered;
- with regards to weighting, question format needs to be considered; and
- marking guidelines need to be considered.

SCIENCE

The Physical Science evaluators for the *2004 Matric Research* adopted the levels of difficulty prescribed in the 2002 Department of Education *Guideline document for Physical Science* (Department of Education 2002). This document prescribed a weighting according to a 'taxonomy of cognitive level', reproduced in Table 8 below.

Table 8: Taxonomy of cognitive levels used by Physical Science evaluators, *2004 Matric Research*

	Approximate HG		Approximate SG	
Cognitive level	**%**	**Marks**	**%**	**Marks**
A. Knowledge, recall, low demand	20	40	30	45
B. Comprehension, routine exercise	40	80	55	83
C. Application, problems, explanation	30	60	15	22
D. Higher abilities, "challenge level"	10	20	0	0
Total	**100**	**200**	**100**	**150**

Source: Department of Education, 2002

The practitioner evaluators in the *2005 School/College Comparison* developed their own four-level grid.

Table 9: Taxonomy used by Physical Science practitioner evaluators, *2005 School/College Comparison*

Category	Cognitive level	Approximate %
1	Reproducing type questions	40 marks
2	Application type questions	25 marks
3	Analyzing type questions	20 marks
4	Evaluation type questions	15 marks

(Chapman, Karsan et al. 2005)

Both taxonomies enabled evaluators to reach clear judgements about the question papers. In some instances, as predicted, the evaluators from the *2005 School/College Comparison*, for example, found that over 70% of the questions in the senior certificate Standard Grade Physical Science examinations required knowledge and understanding of the principles in order to be answered, about 90% of the questions in the Higher Grade papers required the same, but nearly 50% of the vocational Engineering Science N3 could be answered just through knowledge of procedures, that is memorization, repetition and/or own experience. On the other hand, evaluators in the *2004 Matric Research* found that the 2004 Chemistry paper was in fact more difficult than papers of previous years—contrary to expectations. In other words, the tools were found to be useful in focusing the judgements and analysis of the evaluators.

The higher education evaluators from the *2005 School/College Comparison* developed a slightly more sophisticated taxonomy, as shown in Table 10 below. It is more sophisticated because it recognizes that different types of cognitive abilities can be tested at different levels—memorization questions are not necessarily easy, and problem solving questions are not necessarily difficult.

Table 10: Categories and levels of cognitive demand used by Physical Science higher education evaluators, *2005 School/College Comparison*

Category	Level	Descriptions	Examples
Factual recall/rote	Simple (1)	State a simple law or equation	State Newton's laws etc.
	Medium (2)	Recall complex content	Process for lab preparation of chemical compounds; testing for presence of different chemicals; inorganic chemical interactions
Under-standing of concept/ principle	Simple (1)	Simple relationships; simple explanations	Relationship between resultant and equilibrant; explain what is meant by …
	Medium (2)	Counter-intuitive relationships; Qualitative proportional reasoning; more complex relationships or explanations	Direction of acceleration for free-fall; effects of changes in circuits; identifying acid-base conjugates, redox pairs etc; simple influences on dynamic equilibrium

continued/

Table 10/continued

Category	Level	Descriptions	Examples
Under-standing of concept/ principle	Challenging (3)	Identify principles which apply in a novel context	Identify all influences on realistic motion; identify isomers of organic compounds; complex influences on dynamic equilibrium
Problem solving	Simple (1)	Simple procedure; plug into formula with only one unknown; no extraneous information; known or practiced context	Given current and resistance, calculate voltage; etc
	Medium (2)	Construction or interpretation of diagrams; problems with 2 or more steps; basic logic-leaps; proportional reasoning; interpretation of table of data	Graphs of motion; force or vector diagrams; concentration or molar calculations; naming of organic compounds; writing and balancing equations for reactions
	Challenging (3)	Complex abstract representation; combination of concepts across sub-fields; complex problems involving insight and logic-leaps; formulating new equations (using all unknowns); problem solving in novel context	Interpret complex graphs; translate between various graphs of motion; combine equations for mechanical energy and motion; combine gravitational and electrostatic forces; complex circuit calculations; combination of various factors influencing equilibrium

(Grussendorff, Makhafola et al. 2005)

The higher education evaluataors argued that the N3 examination contained no questions which probed understanding of concepts or principles. All the questions fell into either the factual recall or problem-solving categories, and the examination contained no questions at level 3 (the challenging level) of cognitive demand, even in the problem-solving category. Instead, it tested mainly application of procedures (level 1). While their final judgements were not substantially different from the practitioner evaluators, this taxonomy may have allowed a more nuanced judgement to be made.

In the *2006 African Comparison*, Science evaluators did not use the Revised Bloom's Taxonomy as requested. Instead, they again developed their own taxonomy, working with some of the terminology in the Revised Bloom's Taxonomy, but also with the key requirements of Physical Science. This is shown in Table 11 below. PS stands for problem solving, and Pr for practical tasks.

Table 11: Revised Bloom's Taxonomy used by Physical Science evaluators, *2006 African Comparison*

	The cognitive process dimension					
The knowledge dimension	**Remember**	**Understand**	**Apply**		**Analyze**	
Factual knowledge	F1 F2	PS1 PS2 PS3	n/a		PS1 PS2 PS3	 Pr3
Conceptual knowledge	Included in →	C1 C2 C3	PS1 PS2 PS3		n/a	
Procedural knowledge	Included in →	n/a	PS1 PS2 PS3	Pr1 Pr2 Pr3	n/a	

(Grussendorff, Gundry, Makhafola et al. 2006)

The various categories chosen for analysis of examination questions are explained in more detail, together with relevant examples, in Table 12 below:

Table 12: Description of categories and levels of cognitive demands

Category	Level	Descriptions	Examples
Remember Factual knowledge (F)	F1	State a simple law or equation; Recognize content in MCQ	State Newton's laws etc.
	F2	Recall complex content	Process for lab preparation of chemical compounds; testing for presence of different chemicals; inorganic chemical interactions
Understand Conceptual knowledge (C)	C1	Simple relationships; simple explanations	Relationship between resultant and equilibrant; explain what is meant by …
	C2	Counter-intuitive relationships; qualitative proportional reasoning; more complex relationships or explanations	Direction of acceleration for free-fall; effects of changes in circuits; identifying acid-base conjugates, redox pairs etc; simple influences on dynamic equilibrium
	C3	Identify principles which apply in a novel context	Identify all influences on realistic motion; identify isomers of organic compounds; complex influences on dynamic equilibrium
Problem solving (PS)	PS1	Simple procedure; plug into formula with only one unknown; no extraneous information; known or practiced context; simple chemical equation	Given current and resistance, calculate voltage; etc
	PS2	Sketch graphs; construction or interpretation of diagrams; problems with two or more steps; basic logic-leaps; proportional reasoning; interpretation of table of data; acid –base or redox equation	Sketch graph of motion or get information from given graph; force or vector diagrams; concentration or molar calculations; naming of organic compounds; writing and balancing equations for reactions
	PS3	Complex abstract representation; combination of concepts across sub-fields; complex problems involving insight and logic-leaps; formulating new equations (using all unknowns); problem solving in novel context	Interpret complex graphs; translate between various graphs of motion; combine equations for mechanical energy and motion; combine gravitational and electrostatic forces; complex circuit calculations; combination of various factors influencing equilibrium
Practical Procedures (Pr)	Pr1	Simple one-step practical procedure	Reading measuring cylinder
	Pr2	More complex practical procedure; plotting a simple graph	Titration; reading Verniers
	Pr3	More complex graphical analysis; drawing inferences from results	Find relationship between two variables by analyzing gradient of graph

(Grussendorff, Gundry et al. 2006)

Evaluators argued that problem solving forms a significant component of Science examinations. So, then in order to solve a problem, one has first to understand the factual knowlege and then analyze it. This, they argued, is usually considered trivial by an examiner (although it may well not be), and no marks are awarded for this part of the process. After analysis, a candidate has to apply a problem-solving procedure and/or conceptual knowledge in order to solve the problem. In other words, solving a single problem can draw on different dimensions of the matrix presented in Table 11.

Some written examinations test knowledge of practical procedures (for example, the procedure for a titration), and so practical procedural knowledge is included in Table 11 and 12. Evaluators argued that generally this does not involve analyzing given information, but may, for example, require the analysis of a graph derived from an experiment—this is a more complex question, and so would be rated as Pr3.

A question which tests understanding of conceptual knowledge presupposes that the candidate remembers the knowledge. Also, in order to apply procedural knowledge, a candidate has to first *remember* the procedure. However (while it is ideal) it is, in fact, not necessary to understand a procedure in order to apply it.

The evaluators argued that if a Science course is taught well, learners should be expected to evaluate and create knowledge during the course, for example evaluating different models of the atom, or creating knowledge by means of an investigation to find out, for example, which shampoo is best. However, they felt that these skills are not tested in traditional Science exams (such as the examinations under consideration). Similarly, while meta-cognitive knowledge could be included in the teaching of a course, it is simply not tested in traditional Science examinations. Evaluators argued that the question remaining for Umalusi is whether or not it should include these categories as something to aim at, or simply abandon them as unrealistic additions to an already complicated tool.

They also argued, in line with the argument of the English evaluators discussed above, that it is often the way in which knowledge is tested which determines whether it counts as factual or conceptual knowledge, not the content on its own. For example, to remember a statement of Newton's Third Law of Motion is simple recall (F1), but to actually understand the concept embodied by this statement is challenging, since it is counter-intuitive (C2).

MATHEMATICS

The Mathematics evaluation team in the *2004 Matric Research* also used a three-point scale to describe level of cognitive demand or challenge level, with level 1 broadly viewed as 'easy', level 2 as 'moderate' and level 3 as 'most challenging'. What is interesting about their taxonomies, however, is that they developed two different sets for standard and Higher Grade—although Higher Grade is obviously generally more difficult than Standard Grade. This suggests that evaluators needed a sense of hierarchy for the specific requirements of a specific course and examination. The two taxonomies are shows in Tables 13 and 14 opposite.

Table 13: Levels of difficulty in Higher Grade Mathematics examinations, *2004 Matric Research*

Higher Grade	
Level	**Descriptor**
1	Recall with familiar or near-familiar (with respect to expected available learning and teaching resources) formats. Manipulative techniques to be used are given and/or clearly visible. Structural features of the mathematical objects are not complex. Items do not contain any distracting features. Contextual situations are easily identifiable and not distracting in terms of candidates having to grapple with "grasping the context". [Context can be extra- or intra-mathematical.] Translation from "context" to desired mathematical objects are simple due to cues, etc.
2	Manipulative techniques to be applied are to be deduced. Requires the construction of some mathematical objects such as functions, formulae. Cues for translation to mathematical format are not immediately visible in problem presentation. Structural features of mathematical objects differ from the formats presented in expected available learning and teaching resources. Intra- and extra-mathematical contexts require limited interpretation.
3	Structural format of mathematical objects deviates substantially from those found in expected available learning and teaching resources. Working with abstract mathematical entity rather than exemplary embodiment of entity required (Example: Working with f(x) rather than with $f(x) = 2x2 - 3x + 1$ required). Intra- and extra-mathematical contexts require sophisticated interpretation.

Table 14: Level of difficulty in Standard Grade Mathematics examinations, *2004 Matric Research*

Standard Grade	
Level	**Descriptor**
1	Recall with familiar (with respect to expected available learning and teaching resources) formats. Manipulative techniques to be used are given and/or clearly visible. Items do not contain any distracting features. Contextual situations are easily identifiable and not distracting in terms of candidates having to grapple with "grasping the context". [Context can be extra- or intra-mathematical.] No translation necessary from "context" to desired mathematical objects. Mostly single step procedures necessary.
2	Manipulative techniques are near familiar and require a single translation to the standard form. Cues for translation to mathematical format are not immediately visible in problem presentation. Multiple use of mathematical constructs and concepts from different areas are not more than two. Structural features of mathematical objects differ slightly from the formats presented in expected available learning and teaching resources.
3	Mathematical objects deviate reasonably from those found in expected available learning and teaching resources. Limited work with abstract mathematical entities rather than exemplary embodiments of the required entities. Intra- and extra-mathematical contexts require limited interpretation.

(Carrim, Julie and Kitto 2004)

The practitioner evaluators in the *2005 School/College Comparison* developed a five-level grid—perhaps in this way encompassing sufficient distinctions to incorporate different types of papers. Table 15 shows the five levels of their grid, as well as the percentage of each level of question in the college (N3) and Standard Grade school examinations respectively.

Table 15: Grid as applied by Mathematics evaluators in *2005 School/College Comparison*

Level no.	Level descriptor	N3	Standard grade
1.	**Knowledge**: recall of content: facts, notation, conventions, formulae, definitions, standard proofs.	3%	9%
2.	**Skills**: simplification of expressions and solution of standard equations; substitution, drawing graphs…	85%	57%
3.	**Understanding**: Translate from one symbolic form to another; interpret what is required; recognise a tendency; generalise; extrapolate.	12%	21%
4.	**Application**: as for level 3, but situation relatively unknown; method of solution not implicit in the question.	0%	13%
5.	**Creative thought**: Requires the analysis of the material into constituent parts and then a synthesis of these parts to form a new whole.	0%	0%

(Kitto, Nkambule et al. 2005)

Evaluators argued that only the Senior Certificate papers had any questions rated higher than level 3. Their allocation of questions to levels suggests that this five-level grid may have been sufficient to incorporate Higher Grade Mathematics as well, had it fallen within the scope of the evaluation.

The higher education Mathematics evaluators were unable to produce a final report, partly due to disagreements within the group. They did, however, comment on the practitioners' report, using grids supplied in syllabus documents—the four levels for N3 Mathematics (reproducing, application, analyzing, and evaluation) and the seven-point scale to differentiate levels of attainment in the new curriculum statement for the new National Senior Certificate. They also pointed out that:

> [i]t is useful to compare both absolute numbers of marks and percentages. An analysis of similar questions in each paper shows that roughly the same marks are given for the same work. This means that much more is being asked of the SG candidates. Also given the ratio of marks to time, SG candidates are expected to work more quickly (150 marks in 3 hours as opposed to 100).
>
> (Brodie 2005, p. 7)

In the *2006 African Comparison*, the Mathematics evaluators attempted to devise a two-dimensional grid involving rows reflecting the four categories specified in the new South African *National Curriculum Statement* in the *Subject Assessment Guidelines*, and columns using the six verbs from the Revised Bloom's Taxonomy. The four categories from the new curriculum are: knowing, performing routine procedures, performing complex procedures and solving problems. The six verbs from the Revised Bloom's Taxonomy are: remember, understand, apply, analyse, evaluate, and create. Most evaluators found this too complex and finally the group settled on reducing the six columns to three: 1: Remember, 2: Understand and apply, 3: Analyse, evaluate and create. The result is shown in Table 16.

Table 16: Revised grid for evaluation of Mathematics examinations, *2006 African Comparison*

	Remember	Understand, apply	Analyze,evaluate & create	Totals
Knowledge				
Routine procedures				
Complex procedures				
Problem solving				
Totals				

(Kitto, Kock, Mthethwa et al. 2006)

The evaluators attempted to place questions and the marks awarded for these questions in the various cells of the grid. After initial difficulty, further time was spent as a group trying to develop a shared understanding of the types of questions that should be placed in the various cells. Despite this discussion and subsequent communication on the evaluation of a specific paper, the evaluators varied considerably in their placement of a large number of questions.

The evaluators argued that a lot would need to be done to develop a better understanding of this instrument. One evaluator suggested working with just one dimension and another agreed that this might be the solution, but the team was too far advanced in their evaluation to make a change at that stage. In the end, the mean of the totals for each row and column of either three or four evaluators was calculated as a way of reaching some sort of consensus.

BIOLOGY

Biology was not included in the *2005 School/College Comparison*, and thus only the 2004 and 2006 research is discussed here.

In the *2004 Matric Research*, the Biology evaluators dealt with the question of challenge or difficulty level in terms of what they termed "a hierarchy of skills". According to the evaluators, a hierarchy of skills is inherent in performance expectations. For example, it is more difficult to write an essay than to write a paragraph, and more difficult to write a sentence than to recall or recognise a one-word answer.

They list other skills tested in Biology examination papers as follows:

- Draw a diagram: Some questions required candidates to draw a diagram from memory; others asked learners to draw a diagram from a micrograph. Drawing is an important skill in Biology.
- Perform a calculation: Some questions required learners to calculate a numerical value.
- Data response: Questions were categorised as data response when they required learners to extract meaning from data given in a table, graph, experiment or diagram. Questions that required learners to recall the labels on a diagram were not classified as data response.
- Scientific inquiry: Some questions required learners to give the aim of an experiment, or design the control, or identify the controlled variables in an experiment. These were considered to be general scientific inquiry questions, important in the methods of experimentation in Biology.
  (Dempster, Khumalo et al. 2004)

They argued that there is a hierarchy implicit in the non-writing skills required in Biology, for example, it is more difficult to analyze data or a scientific experiment than to draw a diagram from memory. They attempted to eliminate data-response questions based on items that regularly appear in textbooks or in past papers.

In the *2006 African Comparison*, evaluators used the Revised Bloom's Taxonomy to analyze the Biology examination papers from each country. Evaluators spent some time in the initial workshop clarifying the definitions of each knowledge category and cognitive category and jointly applying them to one paper. The analysts then worked independently on the paper already analysed, and the remaining papers. Two analysts worked together and debated questions at length before deciding on a final rating.

The evaluators categorized each question in terms of the knowledge and cognitive dimensions defined, and then rated the questions in terms of three levels of challenge: one, two or three. The levels of challenge were not defined in advance, but as the analysis progressed, it became apparent that the level of challenge could be understood in terms of what learners were expected to produce, or in terms of what reasoning they needed to apply to answer the question. Thus one-word answers would be rated as difficulty level one, constructing a sentence or drawing could be level two, while an extended answer or multi-step reasoning process would be evaluated as level three. These criteria were applied flexibly, because in some cases several steps of reasoning were required to recognize or produce a one-word answer. This point again highlights the need for Umalusi's moderators and other expert evaluators to use tools with care and sensitivity, and see them as providing guidance to experts, as opposed to seeing the tools as rigid and purely scientific.

The findings are extremely interesting, but somewhat complex. While there was some kind of relationship between cognitive operations, types of knowledge, and levels of difficulty, there were also differing levels of difficulty across cognitive operations and types of knowledge. For example, evaluators argued that most of the easy (level one) questions in the South African examinations were in the category *understand conceptual knowledge*, while in the Zambian papers, the easy questions tended to be *recall factual knowledge*.

The question that Umalusi needs to consider is whether or not the effort involved in getting moderators to participate in this type of analysis is really worth it, and whether we will get qualitatively improved moderation.

One final aspect was emphasized by the Biology evaluators: Anderson (2005) clearly indicates that questions that learners have experienced previously would be placed in the factual recall cell of the taxonomy table, rather than in higher level categories. This knowledge was not available to the analysts, although they were all more familiar with the South African examinations than with examinations in the other countries. They pointed out that they are aware that examination questions are recycled in most examining systems, and that students are frequently coached for final examinations by working through previous examination papers. Since they did not know which questions had been used previously, and how much coaching takes place in other countries, they attempted to grade questions as if the learners were encountering the questions for the first time. This observation again raises the need for Umalusi's moderators to use past papers in their work.

HISTORY

History was only considered in the first research project, the *2004 Matric Research.* Evaluators made use of an evaluation grid for history examinations which was part of the syllabus documentation to guide their decisions. The grid which they used, however, was for History essays. The evaluators did not conduct an item-by-item analysis, but did make judgements about the standards of question papers. The guiding grid is shown below in Table 17.

Table 17: Tool used by History evaluators in 2004 Matric Research

Presentation Content	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7
	Has employed analysis and historical explanation. Well planned and structured essay.	Has employed analysis and historical explanation. Well planned and structured essay	Has attempted analysis and historical explanation. Well planned an structured essay	Has attempted analysis and historical explanation. Essay planned and structured to a certain extent.	Applied analysis and historical explanation to a certain extent. Attempted a structure.	Applied analysis and historical explanation occasionally or not at all. No clear structure.	Illogical (to a certain extent). No analysis and historical explanation. No structure.
Level 1 Question has been fully answered. Content selection fully relevant to line of argument.	A+ 47 - 50	A 43 - 46	A- 40 - 42				
Level 2 Question has been answered. Content selection relevant.	A 43 – 46	A- 40 – 42	B+ 38 – 39				
Level 3 Question answered to a great extent. Content adequately covered and relevant.	B+ 38 – 39	B 36 – 37	B- 35	C+ 33 – 34			
Level 4 Question posed is recognisable in answer. Slightly incomplete and mostly relevant.			C+ 33 – 34	C 31 – 32	D+ 28 – 29		
Level 5 Facts do relate to the question but do not answer it. Omissions in coverage (with occasional irrelevance).				C- 30	D 26 – 27	E+ 23 – 24	
Level 6 Facts do relate to the question, but do not answer it. Sparse content (with occasional irrelevance).					D- 25	E 21 – 22	E- 20
Level 7 Question inadequately addressed or not at all. Sparse or inadequate content. Irrelevant to a great extent/ totally irrelevant						E- 20	F 13 – 19 G 07 – 12 H 00 – 06

(Kallaway, Angier and Alexander 2004)

Despite this grid being focused on examination essays and despite the fact that they did not conduct an inventory of questions, the evaluators reached clear and well argued positions on the difficulty level of questions overall.

HOSPITALITY

Hospitality was only included in the second research project, in the *2005 School/College Comparison*. Evaluators developed a three-level tool, with 1 being the most simple. Level 1 was defined as requiring simple recall, often illustrated by the fact that the answer could be found in precisely the asked for form in the textbook. The information required no transformation. Level 2 often required candidates to perform two steps (for example, identify and calculate, compare and contrast and so forth). Level 3 was very rarely found, and comprised mostly application of information (Steyn, Lotz, Venter et al. 2005, p. 9). Examples of the three levels are supplied below.

Level 1:

i) One of the following is NOT classified as a white wine:

A: Sauvignon Blanc

B: Chardonnay

C: Cabernet Sauvignon

D: Rhine Riesling

ii) Name any FOUR types of appetisers

Level 2:

i) Dry cuts of meat can be barded or larded to prevent it (sic) from becoming too dry. Explain the difference between these two terms.

ii) Differentiate between a franchise and a chain group.

Level 3:

South African cuisine is often described as a fusion of colours and flavours, representing all the different ethnic groups which contributed to our heritage. Compile an interesting THREE-course menu which represents this eating style. Write the menu in the answer book. (10 marks)

(Steyn, Lotz et al. 2005)

Evaluators pointed out that these examples could turn out to be inappropriate—for example, learners could have memorized the answer to the level 3 question.

Other factors influencing the quality of examinations

The section below briefly discusses predictability and marking—two additional factors which emerged as clear determinants of the quality of examinations in all three research projects.

PREDICTABILITY

Clearly, for all subjects, predictability was a major determinant of difficulty levels. The Biology evaluators from the 2004 research pointed out, for example, that:

> We attempted to eliminate data response questions based on items that regularly appear in textbooks or in past papers, since these may be answered by recall rather than problem-solving.
>
> (Dempster, Khumalo et al. 2004, p. 5)

The History evaluators of the *2004 Matric Research* pointed out that:

> It seems quite clear from our view of the memos, and from our sampling of a number of the scripts from the 2003 examination, that there is a combination of circumstances here which allows examiners to test knowledge of very predictable topics (often out of date topics or interpretations of history) in ways that make rote-learning pay off. This points to a lack of methodologies and procedures for assessing critical thinking/a tendency to fall back on rote learning/memorization methodologies which could be interpreted as facilitating the increase in the pass rate despite the fact that they deny the overall methodologies prescribed for the examination.
>
> (Kallaway, Angier et al. 2004)

However, Mathematics evaluators from the *2004 Matric Research* pointed out that:

> The kinds of questions that can be asked in a mathematics examination are finite and when an examination has been written for a number of years, the familiarity with these questions increases. Some of the problems date from the very early Egyptian period.
>
> (Carrim, Julie et al. 2004, p. 11)

The Science 2004 report makes the same point:

> In certain sections of the syllabus ('Halogens & Halides' is a good example in Chemistry), there are, to put it bluntly, only so many ways of asking the same question twice. This raises what is potentially a major concern – namely, that certain teachers (and their learners) become adept at playing the 'exam game', thereby exacerbating the tendency within the system of teaching as a kind of 'exam-means to an end'.
>
> (Clark 2004, p. 7)

Nonetheless, predictability is something that moderators should be asked to deal with as far as reasonably possible. To assist them in making judgements, Umalusi should supply the moderators with past papers. It is also worth pointing out that a rigid mark allocation formulation seems likely to lead to increased predictability—this was raised by 2004 and 2005 Science evaluators.

MARKING

Several evaluators picked up problems with marking throughout the research projects. The History evaluators from the *2004 Matric Research* argued that:

> Where more sophisticated questions are asked – either in the SG or the HG – there is often evidence that the examiner (demonstrated by the memo), the moderator and the markers are really not in a position to assess the script in a manner that is sufficiently sophisticated to be able to discriminate between very good, fair or good answers. The marking of such scripts requires a careful in-depth understanding of the material to be covered and an understanding of the debates out of which the questions arise if the grid is to be used to effect. We saw overwhelmingly that marking often simply rewarded the display of content knowledge (memory skills, rote learning, prepared answers) and failed to engage with more sophisticated modes of analysis.
>
> Markers were not always sufficiently discriminating to be able to recognize well-argued and grounded answers.
>
> (Kallaway, Angier et al. 2004, p. 2)

History evaluators also argued that poor marking was leading to a lack of discrimination between mediocre and exceptional candidates. They also raised concerns about the quality of marking memoranda, and about process issues. For example, they argued that "a single session between examiners, moderators and markers in Pretoria prior to the examination does not seem to be sufficient to the task of responding effectively to the diversity of answers expected in this examination format" (Kallaway, Angier et al. 2004, p. 12).

English evaluators from the *2005 School/College Comparison* also found problems with marking:

> A small sample of Business English scripts was examined and the impression was gained of inadequate markers. In one script of Paper 1, question 5 was reached before the student was marked accurately and that question was marked accurately because it was a True / False answer with a quotation as evidence. When candidates answer in the words or almost the words of the passage, they were marked right and given credit. A general principle seems to be that to help markers, the answer is broken down into separate thoughts or concepts and if the candidate synthesises these concepts into a succinct summary the marker does not recognise the 'rightness' of the answer. Being 'good at' English could disadvantage the candidate.
>
> (Allais, Haffejee et al. 2005, p. 6)

Clearly, more attention needs to be paid to improving marking memoranda. However, good memoranda cannot compensate for the lack of ability of markers and perhaps more quality assurance of moderation is required.

Summary of recommendations with regard to evaluating question papers

Umalusi could adopt some of the grids developed by the subject evaluators in the research projects as additional tools for moderators. Alternatively, evaluators could be asked to consider the various grids, and then produce a final grid for Umalusi's use. For other subjects which were not part of any of the evaluations, moderators could be asked to develop a three-scale grid.

Umalusi should try to ensure that the grids which it uses to evaluate examination question papers are contained in the intended curricula.

Umalusi should supply moderators with past examination question papers and explicitly require moderators to check for unreasonable predictability.

Panel moderation should be adopted wherever possible.

In summary: Lessons for Umalusi 7

Umalusi is trying to regularize, standardize, and systematize. It is clear, however, from these three research projects that we must avoid, at all costs, creating cumbersome, overly procedural systems—a danger which is very real. Judgements about the standard and quality of the intended and examined curricula will always rely on the opinions of those who have proven to be, or who are deemed to be experts.

As such, Umalusi needs to think more about the following points:

- How experts are chosen. This is already difficult in schooling but will be much more difficult in vocational education.
- We need to recognise that we will not, and should not, get a situation where all evaluators agree with each other. Nevertheless, we are trying to develop a system to make their judgements as fair and rigorous as possible, and also a system with which Umalusi can do something with their judgements.

Developing tools is useful and the discussions concerning the evaluation of curricula and examinations show that such tools played an important role in the various research projects that Umalusi has undertaken. However, it is clear that *how* the tools are used by the evaluators is much more important than the tools themselves. The tools seem to have been helpful in providing focus for the evaluators.

Reflecting on the three research projects leads us to conclude that, in evaluating the intended and examined curricula:

- some tools are required;
- they, however, play a limited role;
- we need to take into account the role of experts; and
- perhaps most importantly, Umalusi needs to start thinking about what it will do with the evaluation reports.

Considerable work was done on the reports of subject experts for all three of the research projects. This was partly because of our desire to present overall research reports, which meant we needed some consistency across subject reports. In order to obtain reports for systematic purposes this consistency is desirable. It is important to note how different the reports of the different groups were. In at least one instance, evaluators took issue with the final report produced by Umalusi—in the *2004 Matric Research*, some of the evaluators felt the final synthesized report watered down their findings. While this can perhaps been seen as inevitable in a politically-charged research process, it highlights the issue that Umalusi needs to make sense of expert judgements, but currently lacks its own subject experts that can process such judgements.

History evaluators who worked on the 2004 report wrote a letter to Umalusi in which they pointed out that while there was formal continuity over a long period of time in the examinations systems and structures in South Africa, there was not continuity in practices. Specifically, they argue that:

> [t]he role of SAFCERT was quite intentionally to remove academic and expert advisors from the examining board and from the examination panels to remove critics of the apartheid education system. And this has not been reversed since 1994. The old government was keen to remove the universities from this position of influence – and in doing so it did much to take away an academic influence on curriculum policy and assessment.
>
> (Unpublished letter from History evaluators of *Matric 2004 Research*)

Umalusi needs to think very carefully about what we do with reports when we get them, and the kind of expertise that we have to make judgements about the quality of evaluators' and moderators' judgements. The lack of subject experts both at the Department of Education and at Umalusi is in stark contrast to the other three African countries in Umalusi's research. While this is perhaps unavoidable in a system as complex as the South African education system, the overriding conclusion from the three research reports is that Umalusi must avoid the trap of making judgements overly procedural—of trying to use explicit criteria to cover up our lack of ability to deal with and synthesize expert judgements.

REFERENCES

Allais, J., Haffejee, R., et al. (2005). English First Language: Report for Umalusi research comparing school and college subjects. Pretoria: Umalusi.

Allais, S. (2007). The Rise and Fall of the NQF: A critical analysis of the South African National Qualifications Framework. Thesis type. University of the Witwatersrand

Anderson, L.W. (2005). "Objectives, Evaluation, and the Improvement of Education." Studies in Educational Evaluation 31: 102-113.

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

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

Brodie, K. (2005). Umalusi Level 4 Equivalence Research. Analysis of the first report: mathematics. Pretoria: Umalusi.

Buckland, P. (1981). "Another brick in the wall." South African Outlook 112(1331).

Carrim, I., Julie, C., et al. (2004). Report on an Investigation into Senior Certificate Mathematics Examinations of Various Years offered by the National Education Department and the Independent Examinations Board. Pretoria: Umalusi.

Chapman, G., Karsan, H., et al. (2005). Science Group Report: Umalusi comparison of school and college subjects. Pretoria: Umalusi.

Chisholm, L. (2003). The state of curriculum reform in South Africa: The issue of Curriculum 2005. *State of the Nation: South Africa 2003-2004*. J. Daniel, A. Habib and R. Southall. Pretoria: HSRC Press.

Chisholm, L. (2004). The quality of primary education in South Africa. Unesco: http://portal.unesco.org/education, accessed 24th April 2007.

Clark, J. (2004). An Investigation into the Senior Certificate Physical Science Examinations (1992-2003). Pretoria: Umalusi.

Curriculum 2005 Review Committee (2000). A South African curriculum for the twenty-first century. Report of the Review Committee on Curriculum 2005. (Under the coordination of Linda Chisholm). Pretoria: Presented to the Minister of Education, May 2000.

Dempster, E., Khumalo, G., et al. (2004). Umalusi Senior Certificate project: Biology group report. Biology subject report for Umalusi's research into the standard of Matric from 1992 to 2003. Pretoria: Umalusi.

Donnelly, K. (2005). Benchmarking Australian Primary School Curricula. Canberra: Australian Government: Department of Education, Science and Training.

Drummond, J.H. and Paterson, A.N.M. (1991). "Apartheid and Geography Textbooks." Perspectives in Education 12(2): 65-76.

Ferreira, A., Davis, H., et al. (2006). Curriculum complexities of English in Africa. Report of the English group of Umalusi's project comparing South African syllabuses and examinations with those of three other African countries. Pretoria: Umalusi.

Green, A. (2000). "Converging Paths of Ships Passing in the Night? An 'English' critique of Japanese school reform." Comparative Education 36(4): 417-435.

Grobler, M., Akaloo, V., et al. (2005). Investigation into the equivalence between and standards of vocational (National Senior Certificate) general qualifications (Senior Certificate) and the unit standards based resource pack. Pretoria: Umalusi.

Grussendorff, S., Gundry, D., et al. (2006). Umalusi Report on Equivalence of Physical Science Courses. Report for Umalusi comparing syllabuses and examinations in Ghana, Kenya, South Africa, and Zambia. Pretoria: Umalusi.

Grussendorff, S., Makhafola, M., et al. (2005). Report on Equivalence of NQF Level 4 Science Courses. Pretoria: Umalusi.

Hartshorne, K. (1985). "The state of education in South Africa: some indicators." South African Journal of Science 81.

Howie, S. (2001). *Third International Mathematics and Science Study Repeat (TIMSS-R). Executive Summary*. Cape Town: HSRC Press.

Jaffer, K., McKenna, S., et al. (2005). Higher Education Evaluators' Report on the Equivalence of various NQF level 4 English Curricula. Pretoria: Umalusi.

Kallaway, P. (1988). "Privatisation as an aspect of the educational politics of the new right: critical signposts for understanding shifts in education policy in South Africa during the eighties." British Journal of Education Studies 37(3).

Kallaway, P., Angier, K., et al. (2004). Senior Certificate research evaluation exercise. History subject report for Umalusi's research into the standard of Matric from 1992 to 2003. Pretoria: Umalusi.

Kerr, D. (2000). 'National Goals'. World-class Curriculum Conference Papers, Melbourne: Curriculum Corporation.

Kilpatrick, J., Swafford, J., et al., Eds. (2001). *Adding it up: helping children learn mathematics*. Washington DC: National Academic Press.

Kitto, A., Kock, H.d., et al. (2006). Mathematics Report: Comparing the standards of selected subjects at senior secondary level of three African countries with South Africa. Pretoria: Umalusi.

Kitto, A., Nkambule, H., et al. (2005). Report to Umalusi: Mathematics Group. Practitioner evaluators' report for Umalusi comparison of school and college subjects. Pretoria: Umalusi.

Lodge, T. (1983). *Black politics in South Africa since 1945*. Braamfontein: Raven Press.

Lotz, L., Steyn, M., et al. (2005). Hospitality Studies. Practitioner report for Umalusi comparison of school and college subjects. Pretoria: Umalusi.

Lowry, S. (1995). "A Critique of the History Curriculum in South Africa." Perspectives in Education 16(1): 105-129.

Moloi, M. and Strauss, J. (2005). The SACMEQ II Project in South Africa: A Study of the Conditions of Schooling and the Quality of Education. Harare and Pretoria: SACMEQ and South African Ministry of Education.

REFERENCES continued

National Institute for Educational Research (2003). An International Comparative Study of School Curriculums. Osaka: National Institute for Educational Research.

NECC (1992). Curriculum: Report of the NEPI Curriculum Research Group. A project of the National Education Co-ordinating Committee. Cape Town: Oxford University Press/NECC.

Qualifications and Curriculum Authority (2003). Report on Comparability between GCE and International Baccalaureate Examinations. London: QCA.

Schmidt, W.H., Wang, H.C., et al. (2005). "Curriculum coherence: an examination of US mathematics and science content standards from an international perspective." Curriculum Studies 37(5): 525-559.

Spreen, C.A. (2001). Globalization and Educational Policy Borrowing: Mapping Outcomes-Based Education in South Africa. Thesis type. Columbia University Stevenson, H.W. and Nerison-Low, R. (2003). To Sum It Up: Case Studies of Education in Germany, Japan, and the United States. Case study component of the Third International Mathematics and Science Study (TIMSS): National Institute on Student Achievement, Curriculum, and Assessment: Office of Education Research and Improvement, U.S. Department of Education.

Steyn, M., Lotz, L., et al. (2005). Hospitality studies. Practitioner report for Umalusi comparison of school and college subjects. Pretoria: Umalusi.

Umalusi (2004). Investigation into the Senior Certificate Examination. Pretoria: Umalusi.

Umalusi (2006a). Apples and Oranges? A comparison of school and college subjects. Pretoria: Umalusi.

Umalusi (2006b). Cognitive Challenge: Umalusi's role in monitoring standards of qualifications and curricula. Report of an Umalusi Workshop. Pretoria: Umalusi.

Umalusi (2006c). Conference proceedings. 4th Sub-Regional Conference on Assessment in Education, University of Johannesburg: Umalusi.

Umalusi (2007). Umalusi Technical Report: A comparison of syllabuses and examinations from Ghana, Kenya, South Africa, and Zambia. Pretoria: Umalusi.

Victorian Curriculum and Assessment Authority (2004). Curriculum Victoria: Foundations for the Future. Melbourne: VCAA.

Yeld, N. (2005). "Senior Certificate Stagnation and Attendant Woes: Some Consequences for the National Senior Certificate" in Matric: What is to be done? Pretoria: Umalusi and Council for Higher Education and Training.

Yeld, N., Grobler, M., et al. (2004). Investigation into the examination of English Second/Additional Language of selected examination authorities in the years 1992, 1999, 2001 and 2003. English Second Language subject report for Umalusi's research into the standard of Matric from 1992 to 2003. Pretoria: Umalusi.

Appendix 1

Umalusi's moderation guidelines

Preamble

UMALUSI as a quality assuror and custodian of standards with regard to candidate achievements is always striving to ensure that:

1. the standard of question papers set for the examinations is upheld, and
2. the question papers are fair, valid, reliable and appropriate.

The standard of question papers is dependent broadly on the following aspects:

1. coverage of the curriculum content,
2. the kind of questions included in the question paper, e.g. essay questions, short questions, objective items and short items,
3. the quality of questions set in terms of variety, appropriateness, relevancy, clarity and fairness, and
4. the conceptual constructs of the subject that the paper addresses.

The aspects mentioned above are covered in detail below, and should be used by all external moderators when moderating all the examination question papers.

1. Adherence to Assessment Policies/Guideline Documents

- Is the paper in line with the current policy/guideline documents, e.g. NCS and supporting documents?
- Is there an accompanying analysis grid/ assessment framework that illustrates coverage of different cognitive levels and the percentages thereof?
- Does the paper reflect the prescribed learning outcomes and assessment standards?
- Is the weighting and spread of content of the LOs and ASs appropriate as per the Subject Assessment Guidelines?

2. Content Coverage

- Does the paper adequately cover the LOs and the ASs as prescribed in the policy and guideline documents?
- Are the questions within the broad scope of the Statement?
- Does the paper cover questions of various types e.g. multiple choice questions, paragraph, data response, essay, real-life scenarios and real-life problem solving, etc?
- Does the paper allow for creative responses from candidates?
- Is the weighting and spread of content, LOs and ASs appropriate?
- Are the examples and illustrations suitable, appropriate, relevant and academically correct?
- Is there a correlation between mark allocation, level of difficulty and time allocation?
- Are the assessment standards appropriately linked and integrated?

3. Cognitive skills

- Is there an appropriate distribution in terms of cognitive levels (Bloom's taxonomy or any other taxonomy that may have been used)?
- Are choice questions of an equal level of difficulty?
- Is there the correct distribution of marks according to the norms?
- Does the paper provide opportunities to assess the following
  - reasoning ability,
  - ability to communicate,
  - ability to translate from verbal to symbolic,
  - ability to compare and contrast,
  - ability to see causal relationship,
  - ability to express an argument clearly?
- Are the questions (conceptual abilities) representative of the latest developments in the teaching of this knowledge field?

4. Language and Bias

- Is the subject terminology/data used correctly?
- Is the language register appropriate for the level of the candidate?
- Are there any subtleties in the grammar that might create confusion?
- Does the paper have any evidence of bias in terms of gender issues, race, cultural issues, and provincial and regional bias?
- When passages are used, is the text of appropriate length, and is the level and complexity of the vocabulary appropriate?

5. Predictability

- Are the questions of such a nature that they could be easily spotted or predicted?
- Is there repetition of questions from the past three years' question papers?
- Does the paper contain an appropriate degree of innovation?

6. **Marking memo**
   - Comment on the accuracy of the marking memo.
   - Does it correspond with questions in the question paper?
   - Does the memo make allowance for alternative responses?
   - Does it facilitate marking?
   - Is the marking memo laid out clearly and neatly typed?
   - Is the marking memo complete with mark allocation and mark distribution within the questions?
   - Does the memo reflect the LOs and ASs assessed?
7. **Technical criteria**
   - Is the question paper complete with grid, memorandum, relevant answer sheets and formula sheets/addenda?
   - Did the assessment body submit a file containing a full history of the paper with all drafts, moderators' comments etc. (All these must accompany the paper each time it comes to the external moderator).
   - Does the cover page have all relevant details such as time allocation, name of the subject, and instructions to candidates?
   - Are the instructions to candidates clearly specified and unambiguous?
   - Is the lay out of the paper candidate-friendly?
   - Does the paper have the correct numbering?
   - Are appropriate fonts used throughout the paper?
   - Is mark allocation clearly indicated?
   - Is it possible to complete the paper in the time allocated?
   - Is the mark allocation on the paper the same as that on the memo?
   - Comment on the quality of illustrations, graphs, tables etc. Are these print ready?
   - Does the paper adhere to the format requirements in the Subject Assessment Guideline?
8. **Internal Moderation**
   - Is the Internal Moderator's report included?
   - Is there evidence that the paper has been moderated internally?
   - What is the quality, standard and relevance of input from the Internal Moderator?
   - Is there evidence that the Internal Moderator's recommendations have been effected or addressed?
9. **Overall impression of the paper**
   - Is the question paper fair, valid and reliable?
   - Does the paper as a whole assess the outcomes of the Curriculum Statement?
   - Is the question paper of the appropriate standard?
   - How does the standard of the question paper compare in relation to previous years' question papers?
   - Is there a balance between the assessment of skills, knowledge and values?
   - Is the paper in line with the relevant current policy/ guideline documents?

NOTE TO THE EXTERNAL MODERATOR:

- The aspects mentioned under each criterion must be mentioned in the report you submit on each question paper moderated.
- Provide recommendations for improvement or maintenance of standards.
- Please substantiate your decision to either accept or reject certain questions or the whole question paper.

Appendix 2

Guidelines provided to subject evaluators for the *2004 Matric Research*

1. **Content coverage**
   - Does the exam accurately represent the content specified in the syllabus?
   - Are all the key areas covered?
   - Do the questions consist of constructs that represent important details/issues of the areas?
   - Are the questions clear?
   - Are there appropriate numbers of questions for the key areas in the syllabus?
   - Is the mark allocation appropriate to the relative importance of the different key areas of the syllabus?
   - Are there appropriate numbers of questions for the different levels of ability?
   - Is the distribution of marks appropriate to the level of complexity of the different questions?
2. **Cognitive skills**
   - What conceptual constructs of the subject does the paper deal with?

     e.g. – reasoning ability
     - ability to communicate
     - ability to translate from verbal to symbolic
     - ability to compare and contrast
     - ability to see causal relationship
     - ability to express an argument clearly
   - Are these representative of the best and latest developments in the teaching of this knowledge field?
   - Are the questions challenging and allowing for creative responses from candidates?
   - Are the questions correctly differentiated in terms of standard and Higher Grade?
3. **Language**
   - Is the language clear and simple?
   - Is the correct subject terminology used?
   - Is the appropriate register used?
   - Is bias avoided? (gender, religion, race, culture, region, etc)
   - Are the instructions clear?
4. **Marking memorandum**
   - Is the memorandum correct?
   - Do answers correspond with questions?
   - Are alternative answers provided?
   - Is it easy to use?
5. **Length of the paper**
   - Is the overall time allocated to the paper appropriate?
   - Do the marks correspond with time on task?
6. **Technical criteria**
   - Is the paper well-organised?
   - Does the cover page contain all technical details?
   - Is the layout learner-friendly?
   - Is the numbering correct?
   - Are marks clearly indicated?
   - Is the quality of illustrations, graphs, graphs, tables, etc, good?
7. **Overall impression on the paper**
   - Is the paper, as a whole, fair?
   - Does the paper assess the achievement of the aims of the syllabus?
   - Are there any recommendations that can be made for improving the paper?

Appendix 3

Guidelines provided to evaluators for the *2005 School/College Comparison*

Content specification

INTENDED CURRICULUM

Evaluators were asked to describe the intended curriculum, specified in the syllabus and/or learning programme and/or unit standards. The guiding questions were:

- What are the key important areas in this subject?
- How are the different important areas of the subject weighted in the syllabus/unit standards/outcome statements/course pack?
- Do the syllabuses/unit standards/course packs represent an appropriate selection of the domain for this level?

Evaluators were also asked to draw from their knowledge of the subject if, for example, they felt that the syllabuses neglected what, in their view, was a crucial component of the subject in question, and to document these issues. (Here and elsewhere they were asked to clearly indicate where analysis was based on personal experience, and where it was derived from the data.) It must be noted here that curriculum is always a political and highly contested area. Decisions about what should be taught at a senior secondary level are not straightforward, and there is often intense disagreement among experts within subject areas on issues such as what is appropriate at different levels and which areas should be focused on. Such debate and discussion is important, but not always easy to resolve. For the purpose of this research, the opinions of only four individual higher education experts were sought. This inevitably means that the particular views and preferences of those individuals about the focus areas of their particular subjects, or the broad approach which should be taken in their particular subject, dominate some of the findings. For example, the English evaluators did not focus strongly on oral communication, while the Mathematics evaluators felt that there was too much algebra in the syllabuses. It is quite possible that other higher education experts would argue differently on these points.

Evaluators were asked to make judgements about the relative standard of the various courses, as well as the extent to which they prepare learners for higher education, with regard to specified content.

EXAMINED CURRICULUM

Evaluators were then asked to provide a description of the content areas of each subject in terms of the examined curriculum. They were instructed to look at two aspects: sampling and weighting. With regard to sampling, the guiding questions were:

- Is the exam/assessment an accurate representation of the content and skills specified in the syllabus?
- Are all the important areas covered?
- Do the questions represent important details/issues in these areas?
- Are the questions phrased in a way that reflects the core of the issues?

With regard to weighting, additional elaboration was given, because, again based on the criteria used in the *2004 Investigation into the Standard of the Senior Certificate*, weighting was looked at in two different ways. The first was in accordance with the relative importance of the syllabus area which the question covered. The second represented the difficulty of the question—among comprehension questions there are 'factual,' 'interpretive', 'evaluative' and 'synthesis' types of questions. It was pointed out to the evaluators that the 'evaluative' and 'synthesis' question are much more difficult as they require the learner to use what she/he has learned in order to generate an opinion and to draw implications, and hence one would expect a fair distribution of marks and assessment tasks in correspondence to the type of question (or its level of complexity).

The following guiding questions were given in terms of these two aspects of weighting:

- Is sufficient attention/time/coverage given to the key areas of the subject?
- Are there appropriate numbers of questions for the different key areas in the syllabus?
- Are the questions from which candidates are to choose of equal difficulty levels?
- Are there appropriate numbers of questions for the different levels of ability?
- Is the mark allocation appropriate to the relative importance of the different key areas in the syllabus?
- Is the distribution of marks (mark allocation) or relative importance of the assessment task appropriate to the level of complexity of the different questions?

Evaluators were asked to make judgements about the relative standard of the various courses, as well as the extent to which they prepared learners for higher education, with regard to the extent to which the examinations tested the content specified in the syllabus.

The key concepts and procedures

INTENDED CURRICULUM

Evaluators were asked to describe the key concepts and procedures that the subject entailed in terms of the intended curriculum.

The general guiding question which the evaluators were given was:

- Does the content specification refer to a situation related to everyday life (procedural)?
- Does it include questions of an abstract nature (principled)?

This question was broken down into the following guiding questions:

- Is there an abstract concept or principle or general rule that the learner needs to grasp (principled), or is it a list of steps that they are expected to understand (procedural)?
- Are the learners expected to be able to describe something they can see or do, or are they expected to master an underlying idea or approach?

Evaluators were asked to make judgements about the relative standard of the various courses, as well as the extent to which they prepared learners for higher education, with regard to the key concepts and procedures specified for each course.

Examined curriculum

The evaluators were then asked to describe the key concepts and procedures that each question in the exams/assessments (the examined curriculum) measured and the level at which they were measured.

The following guiding questions were given:

- What does this question require of the learner?
- Does it require understanding of an abstract concept or a principle or a general rule?
- Does it require memorization or repetition? Is it something that requires them to repeat what they have learned in class? Is it something they have to remember which has been rehearsed?
- Are they asked to think on their feet and make a new argument? Is it an answer that they have to think out for themselves? How well do they have to articulate their argument? Do they have to reason?
- Where would they get the answer from (are they asked to draw on their own experience, solve a problem or draw on what they have been told)?
- Is it something that requires them to identify and/or apply a concept or rule?

It was not necessary to respond to these guiding questions individually. The evaluators were asked to use these questions to evaluate the level of difficulty of the questions/tasks. It was suggested that the evaluators use a three-point scale, with one indicating the most basic questions, two indicating average difficulty and three involving fairly sophisticated application of knowledge. However, evaluators developed different scales according to their subjects.

Evaluators were asked to make judgements about the relative standard of the various courses, as well as the extent to which they prepared learners for higher education, with regard to the levels of cognitive challenge in the examinations for each course.

Appendix 4

Tools and criteria for the *2006 African Comparison*

The intended curriculum

Evaluators were requested to conduct an analysis of the curricula of the various courses based on the elements and principles listed below. They were also asked to make recommendations for improving the list of key elements and associated guiding questions for the purpose of future work in evaluating curricula/syllabus documents.

ESSENTIAL PRINCIPLES/ELEMENTS OF CURRICULUM DOCUMENTS

Are the following elements clearly present in the curriculum documents? Elements 1-7 apply to all subjects; 8 and 9 are optional.

1. Aims/purpose/vision/outcomes

The syllabus documents should provide a clear sense of the aims of the subject, in relation to content and cognitive operations.

2. Content coverage (breadth)

The key important areas in the subject should be present, and they should be weighted correctly. The syllabus/curriculum document should represent an appropriate selection of the domain for the level in question. Expert evaluators need to reach agreement on what the key important areas are.

3. Coherence

There should be a clear sense of the overall coherence or organizing principle of the subject present in the syllabus or curriculum document. This could be in terms of the internal principles of the discipline, or another clear principle or theoretical framework through which knowledge is organized in the curriculum. The principle should allow for logical progression and should, at any given time, be the dominant paradigm. For example, in Biology, the principle of evolution is the highest ordering principle which holds all knowledge together, and which provides a framework through which teachers can ensure that learners master the bigger picture of the discipline, and not just isolated sets of content.

4. Sequence and progression

There should be progression and appropriate sequencing between content areas, over the course of the year, and between grades/years of study if applicable. Progression could be in terms of content areas as well as in terms of cognitive demand within content areas.

5. Pacing within and across years (if applicable)

There should be clear information for teachers about what should be covered and by when over the entire programme.

6. Content coverage by cognitive demand (depth)

There should be clear and appropriate specification of necessary/appropriate conceptual principles or procedures within each content area. This should also include guidance for teachers with regards to differentiating between difficulty levels, as well as the depth, scope, and range of the conceptual principles. There should be an appropriate range of cognitive operations demanded of learners (such as remember, understand, apply, analyze, evaluate, create).

Note: in order for evaluators to make a judgement about this category, they could look at exemplars, assessment standards, or mock examinations provided within the syllabus or curriculum documents. However, this question will primarily only be able to be answered on the basis of an analysis of examinations.

7. Assessment specifications

There should be clear information about how the curriculum will be assessed, including principles of assessment, types and weighting of assessment tasks, criteria or rubrics, and exemplars. This could include moderation processes where necessary, especially for practical subjects. This should also include the rubric of difficulty levels for the subject.

8. Practical tasks

These should be clear explanations about the nature of practical tasks, as well as the weighting of these tasks.

9. Other

There could be information on teaching approaches, exemplars of teaching tasks and approaches, as well as comments and guidance on materials and resources, for example, what media: textbooks, electronic, library, and so on.

THE EXAMINED CURRICULUM

Evaluators were asked to provide a description of the content areas of each subject in terms of the examined curriculum. Two categories were to be considered: sampling and weighting. With regard to sampling, the following guiding questions were given:

- Is the examination/assessment an accurate representation of the content and skills specified in the syllabus?
- Are all the important areas covered?
- Do the questions represent important details/issues in these areas?
- Are the questions phrased in a way that reflects the core of the issues?

With regard to weighting, evaluators were to consider the relative importance of the different areas in the syllabus (is sufficient

attention given to key areas?), as well as the types of questions—(is there sufficient distribution of different kinds of questions?). The following guiding questions were given:

- Is sufficient attention/time/coverage given to the key areas of the subject?
- Are there appropriate numbers of questions for the different key areas in the syllabus?
- Are the questions from which candidates are to choose of equal difficulty level?
- Are there appropriate numbers of questions for the different levels of ability?
- Is the mark allocation appropriate to the relative importance of the different key areas in the syllabus?
- Is the distribution of marks (mark allocation) or relative importance of the assessment task appropriate to the level of complexity of the different questions?

Scales to evaluate levels of cognitive challenge in examination papers were developed in two previous research projects and further refined in the workshop of experts. However, in previous research Umalusi had allowed subject evaluators to make judgements based on a simple hierarchy of difficulty, and each group of subject evaluators developed their own hierarchy. The tools which were developed for this research, and which have been proposed as part of Umalusi's moderation systems, attempted to capture levels of difficulty within different cognitive operations. The Revised Bloom's Taxonomy (Anderson and Krathwohl 2001) was suggested as a grid that provided a way of applying different cognitive operations to different types of knowledge. Umalusi then suggested a further division into levels of difficulty. We were aware that the resulting tool would be too cumbersome to use, but felt that it could form the basis of a new and better tool, with subject experts removing categories according to the needs of their subject. The Revised Bloom's Taxonomy divided into levels of difficulty is shown below in
Table 1.

The taxonomy differentiates between different cognitive operations, as well as different forms of knowledge. These are then broken into three levels of difficulty. Previous research projects have highlighted the importance of distinguishing between levels of difficulty within different cognitive operations—in other words, it is not good enough to classify memorization as easy, as there could be easy and difficult memorization questions in a paper. Similarly, analyzing or understanding could be tested at different levels of difficulty. Thus, for each cognitive operation, there should be three levels of difficulty.

We anticipated that it would be unlikely that evaluators would feel that all the categories were appropriate for their subject area. In addition, we wanted to ensure the development of a tool that would be easy to use. As such, subject experts were asked to start by refining the tool as appropriate for their subject. This could include, say, combining 'apply and analyze', or 'analyze and evaluate'. It would also include developing examples of what these mean at each level of difficulty for the subject at hand.

Table 1: Revised Bloom's Taxonomy

	The cognitive process dimension					
The Knowledge Dimension	**1. Remember**	**2. Understand.**	**3. Apply**	**4. Analyse**	**5. Evaluate**	**6. Create**
A. Factual knowledge	1	1	1	1	1	1
	2	2	2	2	2	2
	3	3	3	3	3	3
B. Conceptual knowledge	1	1	1	1	1	1
	2	2	2	2	2	2
	3	3	3	3	3	3
C. Procedural knowledge	1	1	1	1	1	1
	2	2	2	2	2	2
	3	3	3	3	3	3
D. Metacognitive knowledge	1	1	1	1	1	1
	2	2	2	2	2	2
	3	3	3	3	3	3

The numbers one to three represent levels of cognitive challenge within each possible cognitive process and knowledge dimension item. This table, with the six categories in the cognitive process dimension, and four categories in the knowledge dimension, was to be used as the starting point for subject experts to develop a subject specific grid that differentiates three levels of cognitive challenge within different cognitive operations.

Appendix 5

Umalusi's guidelines for the evaluation of curriculum statements

PREAMBLE

Judgements about the quality of curriculum statements must be made by groups of appropriately selected experts in the respective subject area. Judgements should be made holistically, based on the experts' knowledge of the subject area. A balance between breadth and depth is necessary in any curriculum and is important, because a curriculum can attempt to cover too much ground, and as such, fail to deal with any topic in any depth, or can cover too few topics in a lot of depth, and fail to build a broad base of skills and knowledge. A good intended curriculum should be focused on what is agreed to be essential (rather than trying to cover everything), specific, manageable for both teachers and students in the time available, focused on conceptual development (rather than long lists of content), sequenced on the basis of evidence (rather than tradition), and supported by shared teacher understanding of what performance 'at the expected outcome or standard' looks like (Kerr 2000).

The guidelines below are not a rigid list of items to be checked by evaluators, but a set of categories to assist evaluators in reaching a considered judgement. Evaluators should bear in mind that the primary audience for curriculum documents consists of teachers, textbook writers and examiners, and should accordingly judge how useful the documents are likely to be for these three groups of people. Curriculum statements should be "clear, succinct, unambiguous, measurable, and based on essential learning as represented by the subject disciplines" (Donnelly 2005, p. 8).

The team of evaluators should consist of no more than five people and should include subject experts, subject education experts, practitioners (preferably a high-performing, experienced teacher and a more ordinary one), and, where needed in vocational subjects, an expert drawn from a professional body or a person with expertise in the specific business sector with an awareness of learning needs. Evaluators should attempt to reach agreement wherever possible; knowing that there is contestation within most disciplines and areas of study about what is essential and appropriate at different areas of study.Evaluators should consider whether the curriculum statement in question would initiate teaching and learning that would prepare learners for further learning at a higher level, in the specific subject under study if applicable, but also whether the curriculum statement, in general, builds the types of skills and abilities needed for further study.

Evaluators may consider examples of internationally well-regarded syllabuses or textbooks that are regarded as exemplary for their particular discipline or subject area as a guiding framework, but should allow for appropriate differences depending on the aims and context of a particular curriculum statement.

Where a new curriculum statement has been developed and is being evaluated by Umalusi, evaluators should be given an exemplar of what the external examination for the subject is likely to look like. Where Umalusi is evaluating already existing curricula, examinations should be considered as a way of judging the likely coverage of the intended curriculum as well as the level of cognitive challenge that is expected from learners. Evaluators should use the categories below as headings for their report where possible and appropriate, preceded by an overall judgement about the curriculum statement. Reports should be succinct and clear.

ESSENTIAL ELEMENTS FOR CURRICULUM DOCUMENTS

1. Content specification and coverage (breadth and depth)

The key important content and conceptual areas must be guided by the discipline or knowledge area, taking into consideration the purpose of the curriculum. The key content and conceptual areas must be weighted in ways that are appropriate for the subject at the level in question, and support meeting the aims and purpose of the syllabus. The description of content should be sufficient to foster teaching of the necessary depth, and yet remain a manageable guideline.There should be clear and appropriate specification of necessary/appropriate conceptual principles or procedures within each content area. This should also include guidance for teachers with regards to differentiating between difficulty levels within the discipline, as well as the levels of difficulty associated with the suggested assessment activities.

There should be an appropriate range of cognitive operations demanded of learners (such as remember, understand, apply, analyze, evaluate, create). Explanations should be supplemented by exemplars of tasks/activities which cover the full range of cognitive operations and difficulty levels for each year of study. These exemplars should not be limited to formal written assessment tasks but should include informal classroom activities.

In language subjects, the syllabus should indicate the types of texts to be covered and prescribe a minimum number of core texts (e.g. novels, plays, poetry, etc.) for each grade/year of study.

2. Organizing principle and coherence

The organizing principle of the discipline or subject determines the emphasis and coherence of the curriculum, and must be clearly stated in the syllabus or curriculum document. The internal principles of the discipline or its theoretical framework should allow for logical progression and should, at any given time, be the

dominant paradigm. A clear sense of overall coherence emerges from the choice of topics, and their ordered connectedness to that organizing principle, as well as the way in which these are assessed. The coherence within the curriculum needs to mirror the discipline itself.

3. Sequence, progression, and pacing

The organizing principle of the subject should allow for logical progression and appropriate sequencing between skills and content areas—over the course of the year, and across grades/years of study. Progression should be evident in content and skills areas, as well as in increasing levels of cognitive demand. There should be clear suggestions about how much time should reasonably be allocated to the various parts of the curriculum, as this would help teachers and examiners identify the relative weighting given to the various parts of the curriculum. The documents must also state how far along the curriculum teachers/learners need to be by the end of each grade/year of study.

4. Aims/purpose/vision/outcomes

The syllabus documents should provide a clear sense of the aims of the subject, in relation to content, skills, and cognitive operations. The stated aims of the syllabus should underpin the design of the syllabus/curriculum as a whole; that is, all aspects of the syllabus/curriculum should work towards enabling the achievement of the aims/outcomes. The aims should not merely function as philosophical statements, and should take account of the broad social context in which the syllabus is to be enacted and assessed. The aims need to be understandable to those using the curriculum.

This section of the curriculum should also show the articulation of the current curriculum with other levels of the education system.

5. Pedagogy and methodology

Suggested or desired teaching approach(es) to the subject domain should be explicitly stated; so, in language, for example, whether a communicative teaching approach or text-based language teaching approach underlies the curriculum. The pedagogy should align with the stated purpose of the curriculum and be appropriate for the likely contexts in which the curriculum is to be used. Clear guidelines are needed to explain what methodologies should be used to meet the pedagogical principles.

6. Assessment guidance

There should be clear information in the curriculum about both internal assessment (continuous and summative) and external assessment (exit-level summative assessment) and their respective weighting. This applies also if there is a practical element as a part of the curriculum.

Principles of assessment, different types of assessment, and criteria and rubrics designed to assess the range of difficulty levels should be clearly outlined. Guidance should be provided as to how many of the different assessment types should be used in each grade/year of study. The weighting of these recommended assessments should agree with the weighting accorded to the different content areas in the curriculum. Assessment guidance should not be so specific as to limit the teaching/learning process. Exemplars of exit-level summative assessments must be provided to indicate the expected level of teaching and learning associated with the curriculum.

7. Practical tasks

Where applicable, the curriculum must provide a clear indication of the nature of the practical tasks associated with the discipline, and should include a diverse range of methods suited to the subject. The weighting of the practical component in the curriculum and in the assessment must be clearly indicated.

8. Provision and packaging of curriculum documents/ syllabus

Curriculum statements or syllabuses must be user friendly (in terms of layout, design and language) and must provide teachers with just the necessary scaffolding to be able to make the best use possible of the information provided. Curricula also need to be freely available to teachers and updates or amendments must be timeously provided.

Although the syllabus/curriculum is likely to comprise multiple documents, the total number of documents should be kept to a minimum; the relationship between the various documents should be clearly outlined; and the focus and purpose of each document should be explicitly stated. The language of the documents should be unambiguous and jargon-free.

9. Processes, designers, and groups consulted in the preparation of the curriculum documents/syllabus

Curriculum statements should be accompanied by a list of the designers of the curriculum, as well as a description of the processes followed in the curriculum design, and of the consultations which took place.

The group that designs the curriculum should include sufficient representation, as required by the subject, from some or all of the following stakeholders: the relevant profession(s), higher education experts, government experts and teachers.

Where possible and necessary, the curriculum should be internationally benchmarked (for example, through providing evidence that a curriculum is clearly equivalent to its counter part in other countries).

10. Other

Curricula may include other elements not covered by the present framework, and it is the designers' prerogative to do so. An example might be that the curriculum makes reference to suitable learning materials and other supportive resources which are aligned to the purpose, rationale and framework set out in the curriculum. Evaluators of such a curriculum should evaluate these additional elements to determine whether or not they add value to the curriculum, and if so, how.

Appendix 6

Possible additions to moderator guidelines

The questions in italics are suggested additions by the English evaluators for *Umalusi's 2006 African Comparison.* Umalusi's unit responsible for the Quality Assurance of Assessment could consider building them into the moderation guidelines.

Sampling—content and skills coverage

1. Is the examination an adequate representation of the **content** specified in the syllabus?
2. Is the examination an adequate representation of the **skills** specified in the syllabus?
3. *Are the **texts selected** an adequate range (register, genre, topic, etc.) and of appropriate levels of difficulty (register, length, conceptual & stylistic density, etc.)?*

Weighting and types of questions

1. Is sufficient attention/time/coverage given to the key content and skills areas of the subject? Are there appropriate numbers of questions for different key content and skills areas in the syllabus? Is the mark allocation appropriate to the relative importance of the different key content and skills areas on the syllabus?
2. Are there appropriate numbers of questions for the different levels of ability? Is the distribution of marks (mark allocation) or relative importance of the assessment task appropriate to the level of complexity of the different questions?
3. *Are the question-answer formats suitable for the content or skills they are assessing?*
4. Are the questions from which candidates are to choose of equal difficulty level?
5. *Are the marking guidelines provided an accurate reflection of the level of demand of the question? Are they sufficient to enable markers to differentiate between answers of different quality?*

What does the question require of the candidate?

1. Does it require memorization or repetition? Is it something that requires them to repeat what they have learnt in class? Is it something they have to remember which has been learnt?
2. Does it require understanding of an abstract concept, or a principle, or a general rule?
3. Is it something that requires them to identify a concept, or rule and/or apply it?
4. Where would they get the answer from? (Are they asked to draw on their own experience, solve a problem, or draw on what they have been told?)
5. Are they asked to think on their feet and make a new argument? Is it an answer that they have to think out for themselves? How well do they have to articulate their argument? Do they have to reason?

The English evaluators argued that the revisions to the guiding questions were useful because they gave an indication of what to read off a completed grid. If the vertical axis has been used to represent content and skills coverage, then it can be used to answer the sampling questions, using the spread and the mark allocation. If the horizontal dimension has been used to capture the type of cognitive process and level of demand, then it can be used, together with the mark allocation, to answer the questions on weighting. The only changes made to the original Umalusi questions have been to add three questions, group related questions together, and omit one or two superfluous questions, with the intention of streamlining this tool. The lengthy list of questions tended to encourage a lock-step question-answer format which did not seem to elicit a clear sense of the different papers.

They also stressed, as discussed above, that it is not possible to evaluate an examination paper without having recourse to the syllabus documentation itself. It is only through a close consideration of the alignment between the syllabus aims, the syllabus focus and the examinations themselves that a clear sense of the overall coherence of a syllabus can be effectively obtained.