The Testing Company

"Too often today's tests measure yesterday's skills with yesterday's testing technologies"

What's holding things up?

David Gleave

david@thetestingcompany.com

So what *is* holding things up?

It is more than five years since the Web-Based Education Commission to the President and Congress made the highly critical and very telling comment: "*Too often today's tests measure yesterday's skills with yesterday's testing technologies*". How much has really changed in those five years? If we consider the external tests and public examinations that control the standards and dominate the pedagogy in our national education systems, the answer would have to be: *very little indeed*. Our school children's skills, aptitudes and achievements continue to be judged on the basis of results from the same old pencil and paper tests, made up largely of multiple choice or 'essay-type' questions, sat in generally uncomfortable conditions and on dates dictated not by the readiness of individual students to take them but purely by the demands of administrative convenience.

This lack of progress *matters* because education systems are driven to a considerable extent by the demands of the tests and examinations that are used to police the curriculum, control key rights of passage and bestow qualifications. For example:

- Primary Leaving tests in many countries grant rights to enter the best schools, but only to the chosen few.
- Middle School exams can be used to select students for academic or vocational programs that will determine their future prospects of employment and higher education.
- Graduation exams at Grades 10 and 12, or tests used to screen for college entrance, are often seen as a (perhaps *the*) deciding factor in the future of young people. They control access to the jobs market as well as higher education – in fact what teachers in some countries refer to chillingly as the students' "*life chances*".

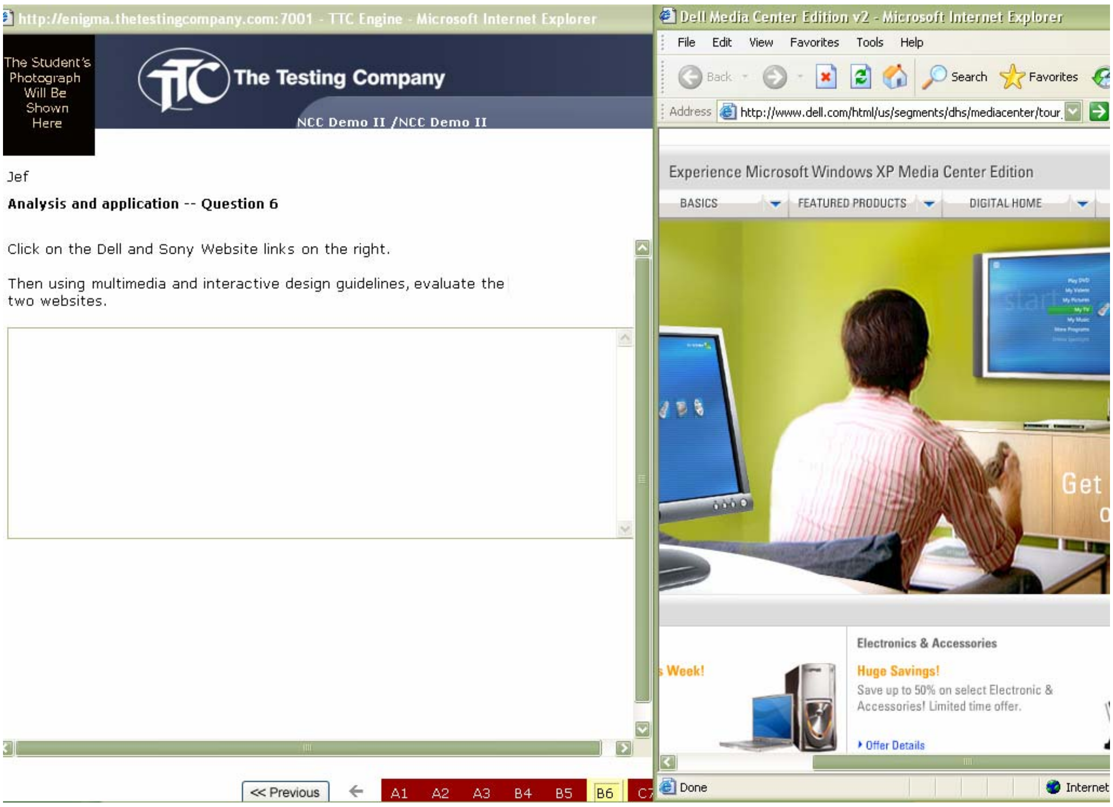
Summative tests like these are expensive to run and can be very disruptive, distracting teachers and students from their normal work for substantial periods. Furthermore they cause a great deal of unreasonable stress and anxiety among school children at several key points in their education.

It also *matters* because educational assessment is ideally positioned to be the quality control mechanism in the education system and such a mechanism is important when every country devotes a substantial proportion of its total resources to education – in money terms between 5 and 10% of the GNP, which makes education the second biggest item in most countries' annual expenditure after defense. Governments and peoples have a right to expect that such huge sums of public money are being spent wisely and efficiently. But, as the Web-Based Education Commission pointed out to the President and Congress, our tests are far too frequently designed to test skills that are out of date or just plain inappropriate to the needs of modern education. Therefore they are not just poor instruments for measuring the quality of national education, they can exert a negative influence through the effect they have on pedagogy and act as an impediment to essential reforms.

Public examinations have a powerful impact on an education system but at the same time they are often one of its most *conservative* elements. Many examination boards are burdened with the duty of protecting cherished national “standards” that have been maintained from generation to generation. Any hint that these ‘standards’ are being tampered with is likely to cause an outcry among the electorate. Governments know this only too well and will go to great lengths to avoid stirring up an issue that can generate a lot of unpredictable political turbulence. In consequence, the assessment systems that dominate our children’s education can sometimes seem like dinosaurs from another age.

Just consider the multiple choice and ‘essay-type’ questions that are still the most commonly used measuring tools in our official systems of educational assessment. The point was made in *Patients don't present with five choices (Veloski et al. Academic Medicine 1999 May;74(5):539-46)* that multiple choice questions are not a particularly effective way to assess the competence of physicians. But are they a particularly effective way to assess competence in *any* field? Modern education generally claims to be about teaching children to understand issues, develop the skills required to solve problems and be innovative. Problems and issues don’t come with pre-packaged sets of alternative solutions. More importantly, the paradigm shift in education is all about getting students to practice new skills *productively* instead of receiving knowledge *passively*. By offering ready made answers to every question, multiple choice tests seem to be completely out of key with this philosophy.

New technology has opened up many opportunities to improve the quality of assessment, including the creation of a whole new range of techniques that are far more appropriate for testing candidates’ abilities.


http://enigma.thetestingcompany.com:7001 - TTC Engine - Microsoft Internet Explorer
The Student's Photograph Will Be Shown Here
The Testing Company
NCC Demo II /NCC Demo II

Jef

Analysis and application -- Question 6

Click on the Dell and Sony Website links on the right.

Then using multimedia and interactive design guidelines, evaluate the two websites.

<< Previous
A1
A2
A3
B4
B5
B6
C7
Dell Media Center Edition v2 - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Address
http://www.dell.com/html/us/segments/dhs/mediacenter/tour
Experience Microsoft Windows XP Media Center Edition
BASICS
FEATURED PRODUCTS
DIGITAL HOME
Get
Week!
Electronics & Accessories
Huge Savings!
Save up to 50% on select Electronic & Accessories! Limited time offer.
Offer Details
Done
Internet

Test designers now have a much broader palette of question types and can select strictly on the basis of *fitness for purpose*. These days the whole taxonomy of educational skills can and should be tested by requiring students to demonstrate these skills in practice - not *passively* by choosing ready made solutions but *actively* by analyzing data, thinking critically, applying knowledge intelligently and working out creative solutions. Unfortunately the psychometric interpretation of statistical data collected from millions of multiple choice test takers over many decades gives the old system an inertia that seems to be all but irresistible. There are many examples of inertia in education, but inertia in the quality control system is particularly damaging because it reinforces out-of-date pedagogy and obstructs curricula reforms that are essential if we are to hold the attention of our students with programs that are relevant to their needs.

It is often claimed that 'essay-type' questions are far better than multiple choice tests at assessing productive abilities, but how far is this really true in practice? More often than not, aren't they just an anachronistic way of assessing skills they are completed unsuited for? The following example is taken from the examination for a business qualification provided by an international testing agency with centres in many countries:

Explain how slides can enhance a presentation. Outline the design principles that should be borne in mind when creating slides and describe how data can be displayed in graphical form.

The objective for this question listed in the examination syllabus is: '*to assess the ability to create a well designed business presentation*'. Well if that's what's required, surely we should be asking the candidate to create a business presentation and use the appropriate and familiar software tools to do so. An online test can set up very effective tasks for displaying this ability by drawing random matching samples of business information from a pre-populated data-bank to provide the content and making the software available within a secure test player. The assessment task is then *fit* to measure the objective, allowing students to demonstrate their design and software skills properly in an authentic context and using the appropriate media. The outcomes will still have to be expertly marked, but now the correct skills and abilities will be judged instead of essay writing skills that are totally divorced from the stated assessment objective. This essay question is not just inappropriate - it creates several other serious problems. One stems from the instructions. "*...Outline the design principles that should be borne in mind...*" It is much more difficult to clearly explain what is required when the task has been alienated from the real objective. The candidate is being made to jump through hoops to demonstrate skills that belong in a different sphere. Writing a good rubric for this type of question can tax the capabilities even of a native speaker. Decoding the intention presents another steep and gratuitous obstacle for candidates to negotiate, particularly if English is not their first language.

Testing experts tend to worry a lot about inter-rater reliability when they think about essay-type questions in test papers. The reports prepared by chief examiners responsible for this kind of assessment identify problems that are far more fundamental than this. Take these comments from a Report prepared for an international examination board whose qualifications are recognized by governments, universities and employers in many parts of the world:

"Many candidates faced difficulties in understanding the requirements of the question."

"Some candidates' handwriting was totally illegible and hard to decipher. Marks could not be allocated appropriately due to ambiguity in the handwriting."

"In general, the candidates are not proficient in the language. As such they are often unable to express themselves well enough to show their understanding of the material."

These are damning comments indeed when the tests concerned were not meant to be assessing either handwriting or English language proficiency.

The problem of inter-rater reliability is very significant and certainly relevant to the subject of this Paper. CBT technology immediately overcomes three of the most basic causes of examiner error commonly found in essay marking: illegibility, bias towards styles of handwriting, and clerical mistakes in recording marks. In the UK, the main thrust of online test development seems to be concerned with scanning written answers and moving script images around, which fails to resolve the first two problems neither of which are minor. Even the Imperial Civil Service examinations of ancient China (from which public examinations in the UK and thus many Commonwealth countries can trace their origins) took the trouble to rewrite answers in a standard script before they were marked to avoid these causes of error.

The development of software for scoring essays automatically seems an equal distraction from more promising avenues of research. Claims that software can mark essays more reliably than human experts certainly shouldn't be accepted at face value. In the first place, there is nothing new in the idea that two markers will give the same essay widely differing scores. Testing agencies have refined procedures for standardizing expert scoring over a very long period of time and have found some reasonably acceptable solutions that can be greatly enhanced by online technology. In the second place e-rating software packages have had it easy until now, undergoing trials in situations where students have no incentive to try foxing the system. Software creators seriously underrate the proficiency of examination cheats, and the skills of human examiners to detect their activities, if they think they can outwit the former and replace the latter with a machine. E-raters haven't risked exposing themselves to the determination of expert essay-test takers in a high stakes context where they really need to get good results. Some enterprising candidates regularly memorize a complete range of well turned out essays written by friends or downloaded from the Internet. The kind of software that bases its judgments on measures of vocabulary level, grammatical sophistication and complexity of structure wouldn't last one round in this environment. Experienced examiners can detect these bogus answers, but e-raters? No, I don't think so.

Does this mean that technology cannot improve on traditional marking and moderation practice - far from it! The whole process of standardizing expertly awarded scores can be carried out a great deal more thoroughly and efficiently online. Essay answers can be 'apportioned' by question or even sub-part of a question and made available for multiple marking online in any and every time zone in the world the instant the candidate hits the End Test button. Online systems alert markers by email to the fact that they have tasks on their marking page, display marking schemes and highlight key words in the answers, provide marking tools, carry out any required calculations and collect raw marks without making any mistakes. The same script can be marked by any number of expert markers at the same time and the system will automatically pass all discrepant marking up the hierarchy to be checked and amended by a higher authority. By the time expert marks are confirmed as final, they will have been far more extensively scrutinized than would have been possible in any traditional system. The resulting scores will be many times more reliable and available in a fraction of the time. Furthermore all the answers, the whole process of marking and moderation, and the detailed history of every mark can be permanently stored for future reference. The system never loses either scripts or data.

Above all, technology provides the opportunities and makes the time and cost gains necessary to ensure that sufficient expert manpower can be efficiently deployed to check and standardize expert marking properly while staying within a reasonable budget. This completely obviates the highly dubious practice of 'examiner scaling' carried out by traditional examination boards in certain countries and education systems. It would probably create an uproar if candidates knew that their marks can be scaled up and down significantly without their answers even being seen a second time. But people have no real understanding of how marks are 'processed' and final results are decided in public examinations. The UK system has been described by the government's chief regulator of examinations as *"a cottage industry"* and the arcane procedures used for arriving at final grades are so impenetrable that a frenzied argument breaks out every time results are released about whether or not the 'standards' have been inflated. In the US, confusion about the meaning of test results is often linked to accusations that *"the private testing companies that control standardized testing operate behind closed doors with little to no public accountability."* (*Barbara Milner, Testing Companies Mine for Gold, Rethinking Schools, Winter 2004/5)* This echoes the commonly expressed view in UK that any queries made to an examination board will be *"met with a wall of bureaucracy and secrecy ... We need to open up the exam boards so that their operation is much more transparent. Only in this way will we stand any chance of restoring public confidence in our battered examinations system."* (*The Observer, Sunday September 22, 2002).* However, in spite of the suspicions they arouse, it is not secrecy that shrouds the workings of educational assessment agencies. Their extraordinarily abstruse procedures and impenetrable jargon combine to create a far more effective barrier to public understanding than secrecy ever could.

Responding to the findings of a recent commission looking into educational standards in the USA, Professor Sam Wineberg of Stanford University lays the blame for poor history teaching in American schools squarely at the door of multiple choice tests. His argument is that, in a norm-referenced system, objective test questions are selected according to their statistical ability to spread candidates evenly across the bell-curve without much reference to whether they test anything worth teaching. And his conclusion is that "*...as long as historians roll over and play dead in front of number-wielding psychometricians, we can have all the blue-ribbon commissions in the world but the results will be the same.*" (*The Journal of American History, March 2004).* His argument may not cut much ice with testing companies that have been shrugging off arguments about lousy test questions for years. But he puts his finger right on one of the main sources of public discontent with our educational assessment systems and that is norm-referencing. The end users of test results often complain loudly that they don't know what the results of a test *mean*, in practical terms, about an individual candidate's knowledge and ability in the subject of the test. But the plain fact is a norm-referenced test can't tell us anything about these things. It can only tell us where a particular candidate shows up on the bell curve and therefore how his/her results compare with those of all the other candidates who took the test, and to a certain extent how they compare with the results of previous generations of candidates who took the same kind of test. Ask any assessment agency with norm-referenced tests to list the common things that all candidates getting a particular result in a particular test can do and you will get a very stony response. They have absolutely no idea. Norm-referenced tests are not about judging what students can or can't do. They are only about placing students accurately on a bell curve. A lot of people – including university admissions tutors, employers, teachers, parents, and the students themselves - are not happy with this because they don't understand it. It doesn't tell them any of the things they want to know.

British exam boards saw the writing on this particular wall some time ago and have been claiming that their system is a hybrid. But this should be taken with a large pinch of salt. They can't list the common things all candidates getting a Grade C in, say, Math at A level can do. So their exams certainly aren't criterion-referenced. On the other hand there are very special reasons why apologists for the UK GCSE level and Advanced level examinations claim that their system is a hybrid. The situation in England and Wales is very unusual, if not unique. These public examinations, offered in all the main subjects of the school curriculum, make up the graduation qualifications for students completing their Grade 10 and Grade 12+ courses. The results of the examinations for each subject are reported against a common set of grades for each level. The extraordinary feature of the system is that schools can choose to enter their students for different tests offered by different examining bodies in the same subjects *and the results are reported on a common grading scale.* This makes for a very complex mix of candidate populations because schools don't stick with a particular board; they can shop around subject by subject and from exam session to exam session. Comparing grading standards across boards by statistical methods is a nightmare and yet results have to be reported on the common scale and so they must be exactly equivalent. The boards are forced to take the statistics as a rough guide and then try to use expert judgment to come to the final decisions. The problem is that the tests are not designed and built to assess mastery and so making decisions on the basis of total scores looks like a pretty hit and miss affair. Major discrepancies in the proportions of candidates judged to be of a particular grade between one board and another then have to be explained by reference to differences in the nature and quality of the candidate populations. However, these are very difficult to quantify and in any case boards have only the vaguest ideas about the populations of candidates they are testing. *"Differences between boards of this kind are usually attributed to the sorts of schools which use them. It is said, for example, that OCR exams tend to be taken by a higher proportion of independent and selective schools, accounting for their higher grades. WJEC thinks it is used mainly by schools in more affluent areas." (How exam results vary between boards, BBC News, 17th September, 2002).* The problems with Advanced level results in 2002 led to greater control over standards, but the official Report on the debacle pointed to the need to increase *"the use of ICT in the administration and marking of public examinations and eventually in the examining process itself." (Final Report of the Tomlinson Enquiry - December, 2002).*

The title of this Paper is 'What's holding things up?' One reason is that a lot of the early attempts at computer based testing were disappointing to say the least and bad experiences have served to sour perceptions and reinforce conservative attitudes throughout the industry. The main shortcomings stemmed from one or more of the following:

a) insufficient understanding of the real needs of educational assessment;
b) a 'hard-wired', one-system-fits-all approach;
c) failure to appreciate the need to adapt procedures and methodology to take full advantage of the possibilities offered by the technology.

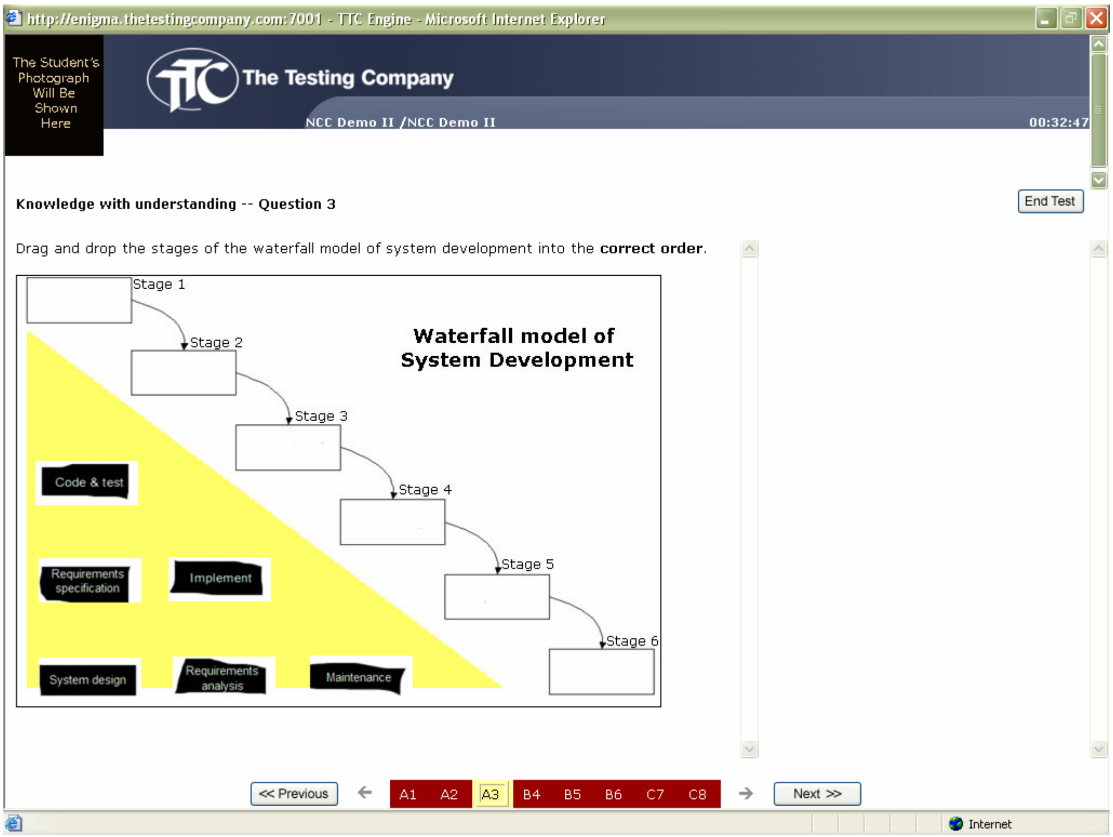
The last of these is still proving to be a major obstacle to progress. Testing Agencies and Ministry of Education examination departments, which control the biggest R&D budgets, often follow procedures that have been used for decades and they are extremely wary of making changes. On the other hand, trying to smash the technology into a shape that will fit the old ways is like asking a film producer to put the text of a novel up on the cinema screen instead of making the film of the book. This is how we end up with multiple choice tests online and huge volumes of essay answers being digitally scanned. Online testing is a highly disruptive technology threatening a very large and very traditional industry. But the fact is that existing public examination regimes are under heavy fire everywhere in the world and boards are finding it increasingly difficult to convince the public that their tests are either fair to students or relevant to the needs of modern education. Maybe this is the *ideal* time for us to respond to market demands and give our customers some of the things they want:

- Tests that genuinely assess the skills our children need instead of reinforcing passive learning.
- Informative results that identify weaknesses and help students to improve instead of scores and grades that convey very little information of any value.

- Standards that can be understood without the need for a degree in psychometrics.
- ‘On time assessment’ when students are ready and not *en masse* exams that encourage phobias and poor performance.
- Reliable data about the efficiency of schools that can be used by governments to improve the quality of education.

Web-based technologies have the power to completely transform educational assessment, not only making it much fairer and many times more valid and reliable but turning it into an integral part of the learning process with very real benefits for every student. Perhaps first of all we should stop treating test candidates like cattle. It is almost incredible that we allow our children to be herded on specially appointed days into vast exam halls where they are tagged like sheep for the slaughter and penned behind little tables for hours on end. No account is taken of whether or not they are *ready* to be assessed, or the extent to which the totally artificial environment might affect their responses, and least of all about all the anxiety they have been made to suffer beforehand. The *main* reason for forcing students to take all their tests at the same time at the end of their courses of study seems to be administrative convenience, because secure pencil and paper tests are the very devil to organize. But it is a poor way to assess students if we place any importance on the need to improve educational efficiency. An online system can randomly generate criterion referenced progress tests that will indicate how well individual students are mastering the learning objectives of the course units they are studying. This information can be made available to teachers immediately so that students with difficulties can be helped and high fliers can be allowed to move ahead. Projects, assignments and other forms of coursework can be uploaded to the system and moderated online. So can many other kinds of file, including digital photographs, videos of personal performances or presentations, and results of other measures such as grade point averages. Together with a complete record of the student’s performance in progress tests, practice tests and end of course summative assessments, all this data provides a comprehensive picture of the student contained in a permanent Student Assessment Record that can be viewed online and used to create a digital portfolio for external users to view.

The *mode* of assessment is clearly very important, but in fact a good online system can substantially improve almost *every* area of the testing process. Take cost and efficiency for example. There are no test papers to print, collate, pack and ship; no answer papers to collect and move around. Item writers fill in computer generated question templates online and tests are created automatically in accordance with pre-set specifications. Answers to objective questions are system scored and free responses are multiple marked online; discrepant expert scores are automatically referred to a higher authority and standardized before being processed. Then there is security. This is much easier to handle when randomly generated tests are created from a central server specifically for individual students to take at set times on recognized PCs in registered and proctored test centres. The results of a Web-based test can be viewed immediately if the questions were objective and the answers were system scored. Even when the test includes expertly scored elements, the results can be available within a few days if marking and standardization have been properly organized. But the greatest and most significant improvements are in the sheer *quality* of assessment. CBT permits the use of an ever-expanding range of question types. Every kind of multiple choice and essay-type question can be supported of course, that goes without saying. But then so many more objective question types are available that can be scored automatically: multiple answers, matching columns, drag and drop, hot spots, completion and limited answer cloze, as well as algorithmic items which the system can replicate many times by simply changing the variables within pre-set ranges.



Knowledge with understanding -- Question 3

Drag and drop the stages of the waterfall model of system development into the **correct order**.

Both system scored and expertly marked free response questions can be accompanied by additional material in the form of text, graphics, animations, audio or video clips, tables of figures, maps, graphs, etc. Case Studies can be created, made up of many data files, providing authentic materials on which extended questions assessing analysis, synthesis and evaluation can be based. Finally, application packages can be launched within the secure test player allowing candidates to use the appropriate software tools to tackle complex problems and engineer solutions. Many teachers resist the idea of online testing because they are under the completely mistaken impression that it can't handle the assessment of 'higher level skills'. Nothing could be further from the truth. There are almost no limits to the kind of skills that Web-based technology can help us to assess very reliably, in the right contexts and using the appropriate tools.

Continuous assessment, which Web-based systems can handle particularly well, offers a very good starting point for Ministries of Education wishing to introduce online testing gradually without causing public alarm, disturbing existing programs, or incurring enormous expense. Online systems are extremely efficient and hence cost-effective. Perhaps most importantly, they overcome the problem of biased marking that has bedeviled so many attempts to use school based assessment as a significant element in national examinations. A reliable series of tests designed to assess the objectives for consecutive units in a course can be randomly generated on demand and scored automatically. This kind of assessment can be managed with only a few PCs in a school, maybe one to each participating class. By monitoring the extent of every individual's mastery of the required learning outcomes as they progress through a course, the results provide teachers with the information they need to remediate problems and ensure the success of all pupils in the class. Where students have access to the Internet in a school library, at home, or even in a computer games shop, they can take on demand practice tests to improve the scores in their assessment records. Many countries fear they don't have sufficient band width available for this kind of testing, but on-demand assessment spreads the load evenly and therefore thinly. Besides which, well designed online tests use very little bandwidth and the normal telephone system should be adequate in all but the most exceptional cases.

The great educational advances promised by the Internet will only be realized when boards responsible for public examinations take the lead by adapting their methods and procedures to take full advantage of the opportunities offered by Web-based assessment systems. Schools, teachers and students respond to the demands of these examinations because they are the official measure of their success or failure. Many governments have equipped their schools with computers and provided them with Internet access without seeing any commensurate improvements in the quality of education. A national system of online continuous assessment would provide an immediate and very powerful incentive for teachers to begin using the technology properly. Regular progress tests can accurately measure student performance in relation to curriculum objectives and other standards and norms. This in turn provides reliable data on which to judge the efficiency of educational provision nationally, regionally, by sector, by school and so on. Governments need reliable national data to accurately estimate returns on educational investment in terms of real improvements at the classroom level and to report such improvements in ways that will be publicly credible. But the most important long term gains will come from getting teachers actively involved in the process and using the technology to upgrade the quality of the learning environment for themselves.