SNAPSHOT

How should we test our students?

5 expert views on assessment

From *World Class: Tackling the Ten Biggest Challenges Facing Schools Today*, edited by David James and Ian Warwick

Routledge
Taylor & Francis Group

Routledge... think about it
www.routledge.com

Introduction: David James and Ian Warwick

For many people, examinations are, literally, the stuff of nightmares. Many adults admit that the bad dreams they have, the ones that wake them at night, sweating, or crying out for a parent, return them to the examinations they took many years ago. They share the same imagery: the examination hall with regimented desks and chairs, the clock ticking down to the final minutes, the prowling invigilators, the beating heart, Poe-like in its intensity, and as you turn the page over and discover you can answer... nothing. Cue: waking up, screaming.

Why do we need to test students? Why persevere with something so traumatising (and imperfect)?

It was only a dream, but testing is very real and very big business. The regularity of such dreams may or may not be an urban myth, but they resonate so much because they connect with us, almost viscerally, and in doing so they help perpetuate the idea that examinations induce more stress among students (not to mention parents and teachers) than any other aspect of schooling. And although some cultures place greater store in examination results, they are deemed to be almost universally necessary, even, as Tim Oates says in this chapter, in 'that apparent educational utopia, Finland', if only to provide information for the next stage of a student's education.

Why do we need to test students? Why persevere with something so traumatising (and imperfect)? Do we retain tests because they keep students focused, providing them with an incentive, and a goal to work towards? If we really do need them can we not assess progress in a more imaginative, 'low stakes' way so that teaching can be liberated from constant monitoring of ever-improving results? Such questions are routinely asked by anyone involved in assessment every year, but solutions rarely follow, and somehow the machinery that sends millions of students each year into those nightmare-inducing rooms grinds on, mostly unchanged. Could it be that it grinds on because it works better than any realistic alternative that could be scaled across national systems? Possibly.

Of course, concerns about the nature of assessment go deeper than simply not liking examinations. For Carolyn Adams and Matt Glanville, 'assessments drive teaching and learning, especially when they are high-stakes university entrance tests'. And for many teachers this is precisely what is wrong with standardised assessment models: 'teaching to the test' is a pejorative term because it suggests a narrowing of ambition, and a utilitarian, reductive approach to learning. To put it another way, if it isn't in the final exam there's no point in teaching it because the only thing that matters is what can be measured.

But how fair is such an approach, both for schools and students? Yes, it might mean the students get the grades they want, but how does that prepare them for life, regardless of whether they want to go on to study subjects at an increasingly higher level? Why are highly selective universities reintroducing their own examinations to better assess academic ability among students? And amid all this debate it is often too easy to forget that something intangible – the sheer joy and love of learning – is neglected because it remains stubbornly untestable. And of course, teaching to the test can often result in dull, repetitive teaching.

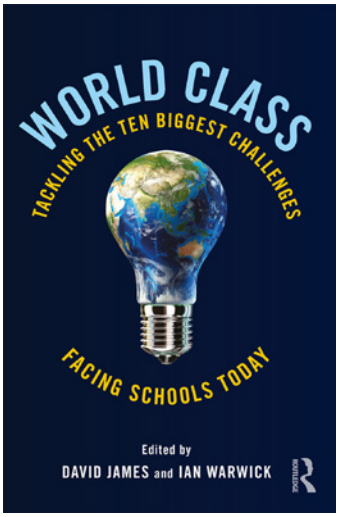
As Dylan Wiliam points out, tests that are essentially intended to assess student progress are also being used for other purposes, including 'what students have learned, what they need to do next, how good their teachers and schools are, and even how schools in on country compare

with those in another'. No wonder they cannot meet all these requirements, and no wonder that one tool is incapable of crafting so many different outcomes.

Perhaps new technology will provide the nuanced approach we have lacked in the past. In this chapter, Andreas Schleicher explores how, in some respects, the future is already here. He writes that 'it is now possible to create coherent multi-layered real-time assessment systems that extend from students to classrooms to schools to regional to national and even international levels'. Such developments, in Schleicher's view, not only assess understanding more accurately, but could inform future policies, again at both local and national levels. For Wiliam, over the next ten years, assessment will change 'beyond recognition', and much for the good, but there are risks, and for Oates, whatever form assessment takes in the future, we must 'meet the challenge of being authentic and accurate'. Placing too much faith in new technology, and in particular in collecting more (and increasingly 'clean') data, could endanger such authenticity, allowing us to lose sight of the student. If that happens then assessment might become, at best, of marginal interest to schools and, at worst, a distorting factor that warps education into something certainly more measurable, but infinitely less valuable.

How should we test our students? 5 expert views on assessment:

1. Carolyn Adams and Matt Glanville
   Director of Strategy Development and Execution; and Head of Assessment Principles and Practice, The International Baccalaureate
2. Tim Oates
   Group Director of Assessment Research and Development, Cambridge Assessment
3. Andreas Schleicher
   Director for Education and Skills, and Special Advisor on Education Policy to the Secretary-General, Organisation for Economics Co-operation and Development (OECD)
4. Bob Sornson
   Acclaimed presenter and best-selling author of Over-Tested and Under-Prepared: Using Competency Based Learning to Transform Our Schoools, and many others
5. Dylan William
   Emeritus Professor of Educational Assessment, University College London



This is an excerpt from:

World Class
Tackling the Ten Biggest Challenges Facing Schools Today

Edited by David James and Ian Warwick

Carolyn Adams and Matt Glanville

1

> What matters is not the absorption and regurgitation either of facts or of pre-digested interpretations of facts, but the development of powers of the mind or ways of thinking which can be applied to new situations and new presentations of facts as they arise.[i]
>
> *– Alec Peterson*

What is best for the student is not always judged based on long-term educational benefits.

Assessments drive teaching and learning, especially where they are high-stakes university entrance tests, which amplifies the importance of the choice of qualification not only for students, but for schools.

The question, 'What is the best qualification for students to take?', means more than that which gives the best chance of a high score for short-term university entrance. Will the qualification stretch and develop students' curiosity? Will it give them skills to succeed in higher education and future life? Or is it just one where they can obtain good marks by careful teaching to the test, a predictable assessment which assesses basic recall of knowledge?

These questions should be asked on a regular basis. Yet educators have stopped asking them as external pressures from national systems make the latter the only possible choice.

The International Baccalaureate's (IB) Diploma Programme has a compulsory core that includes a 4,000-word 'extended essay' on a research question of the student's choice, and a 1,600-word essay on a 'theory of knowledge' topic. The core is worth only three points out of a total of forty-five, but it requires significant effort on the part of students. A 'teaching to the test' approach would place little emphasis on the core as the points allocation is small. But IB research consistently finds that students and teachers believe that the core makes a substantial contribution to students' success at university.

This tension in determining the best for the student is something that we see all around the world. Sometimes, this is explicit, where university entry depends on high scores, and other times more subtly where subjects are chosen to demonstrate particular expertise, as in English language or calculus. Teachers are undeniably trying to maximise the benefit for the student in their chosen qualification, but what is best for the student is not always judged based on long-term educational benefits.

In a rational world, a school should consider both the benefits and disadvantages for the student of each approach. But is the educational climate in some countries removing the element of choice?

The IB encourages teachers to talk about this issue. What are the principles and ethos of the school and how well do they prepare their students for a future which is 'volatile, uncertain, complex and ambiguous (VUCA)'? Only by openly challenging the short-term gains of presenting their students for assessments based largely on knowledge recall and lower-order skills can educators seize the opportunity to consider the purpose of their students' education in a balanced context.

Assessments which require students to think and which demand higher-order cognitive skills are not only for high-achieving students. The world is no longer a place where students who were forced down a rigid disciplinary route can expect to succeed in later life. Today's young people need to make connections between and across subjects. They need to understand and embrace the diversity of disciplines, cultures and languages. They need to understand the nature of knowledge and how to apply it rather than simply learn facts, which are readily available through the cheapest smart phone. Assessments should address those needs to be relevant and educationally worthwhile.

Schools take some immediate risks when they take the 'hard road' of prioritising the educational needs of their students to thrive in a 'VUCA' future over traditional assessments. The risks are that universities, still largely structured in disciplines themselves, might continue to advantage the student with the traditional qualification. National governments might shy away from tolerating a less-objective approach to assessment where more than one response might be accepted as correct. Forward-looking schools might not be able to compete fairly in a 'league table' world.

However, given the promise of connecting assessment more fully to learning, shouldn't we be willing to manage those risks and consider alternative approaches?

2

Tim Oates

Assessment remains heavily contested. For pupils, an assessment in which they might be seen to be failing to meet expectations invokes stress. For teachers, formal assessment continues to be associated with both heavy workload and, where results are used for accountability, their own stress. Even in that apparent educational utopia, Finland, the density of structured testing is far higher than common accounts usually suggest, with teachers concerned about the workload it represents. For parents, formal assessment similarly invokes anxiety, where this represents children falling behind 'expectations' and/or destines them for lower(er)-status routes within a nation's educational arrangements. Over in one 'zone' of the public discourse about assessment, the language invokes ideas of stress, anxiety and workload. These negatives are also linked to disputed fairness, accuracy and intolerance of errors, human and other. This is the zone of 'high-stakes' assessment, where the uses of assessment – in determining routes, in school accountability – causes increasing pressure to build. Over in the 'low-stakes' zone, nations have striven to introduce 'assessment for learning' or worked hard – not always successfully – to ensure that classroom assessment supports and enhances learning, without contributing to increased adverse stress on rather than of pupils. Naturally, the 'high-stakes' or 'low-stakes' nature of assessment is not a feature of the test itself; it derives principally from the uses to which an assessment is put. Almost identical tests rear up as high stakes in some settings, and low stakes in others. Some tests have transmuted from 'low stakes' into 'high stakes' seemingly against the stated wishes of the state and the test designer. The 'phonics check' was introduced in 2012 in schools in England as a means of establishing whether children in the first year of formal schooling have developed an 'acceptable' level of phonic decoding – an important issue for equity and attainment. Designed as useful, practical help to schools, the test rapidly became controversial and strained as it became associated with national analysis and reporting of results. While the phonics check is located at the beginning of education, at the later stages in many

Qualifications are becoming increasingly critical as both higher education and labour markets become more focused in selection.

systems, qualifications are becoming increasingly critical as both higher education and labour markets become more focused in selection. Stakes seem to be increasing in so many segments of assessment.

The social and economic uses of assessment of course exert pressure on the technical characteristics of assessment: precision, fairness, dependability. As the assessments come to matter more, so the pressure on accuracy increases. The stresses thus are felt not only by pupils, parents and teachers, but by the producers of assessments, too. The quality of questions, the quality of marking, the avoidance of practical errors in administration and processing all become a greater focus of attention. Assessment bodies have responded with on-line marking, automated 'real-time' monitoring of marking, all necessitating massive investment in background systems.

Meanwhile, talk of 'the future of assessment' seems to be dominated by discussion of the unfettered promises of new technology: on-demand tests, adaptive testing, 'high-density' formative assessment, 'flipped learning'. All of this feels like two 'futures' – the first linked to the 'high-stakes' present – a dystopian, grimy and harsh urban landscape, while the second seems a utopian idyll full of bucolic promise. Assessment theory and practice appears fraught with this intense opposition. And psychological theory says it's ultimately creative to be depressed (future 1?), while elsewhere in the canon we find it's essential to dream if we are to determine effective solutions to today's problems (future 2?). This schism seems reinforced by the fact that we have been slow to ensure that assessment specialists are part of the groups developing new assessment platforms and, conversely, to ensure that assessment groups are supported by technology experts able to develop new, powerful systems. But one thing does in fact unite these two apparently phobic realities. And it's this. For all the discussion of the apparently 'mired and oppressive' forms of formal assessment, and the 'liberating, learner-centred' character of low-stakes assessment, all assessment is measurement. Underneath, it must meet the challenge of being authentic and accurate. It's about asking people to do things or answer questions, and then from what they do or say inferring what they know, understand or can do. And whether it's low stakes or high stakes, technology-supported or not, there is a common interest and common good in assessment being accurate and dependable. Forgetting this fundamental would be a huge error.

Andreas Schleicher

3

There has never been a greater opportunity to move the assessment agenda forward from providing signals of what students can do, to actually improving what students can do.

The demands on learners and thus tests are evolving fast. In the past, education was about teaching people something. Now, it's about making sure that students develop a reliable compass and the navigation skills to find their own way through an increasingly uncertain, volatile and ambiguous world. These days, we no longer know exactly how things will unfold, often we are surprised and need to learn from the extraordinary and sometimes we make mistakes along the way. And it will often be the mistakes and failures, when properly understood, that create the context for learning and growth. A generation ago, teachers could expect that what they taught would last for the lifetime of their students. Today, teachers need to prepare students for more rapid economic and social change than ever before, for jobs that

Today, teachers need to prepare students for more rapid economic and social change than ever before.

have not yet been created, to use technologies that have not yet been invented, and to solve social problems that we don't yet know will arise.

The dilemma for educators is that the kind of skills that are easiest to teach and easiest to test, are also the skills that are easiest to digitize, automate and outsource. There is no question that state-of-the-art disciplinary knowledge will always remain the foundation. Innovative or creative people generally have specialized skills in a field of knowledge or a practice. And as much as 'learning to learn' skills are important, we always learn by learning something. But one can solve large parts of today's school tests in seconds with the help of a smartphone. If children are to be smarter than a smartphone, then tests need to look beyond whether students can reproduce what they learned to see, whether they can extrapolate from what they know and use their knowledge creatively in novel situations. Put simply, the world no longer rewards people just for what they know – Google knows everything – but for what they can do with what they know. Future tests should not penalise students for connecting with the web, but encourage them to do so.

Conventionally, our approach to problems in schooling is to break them down into manageable bits and pieces, and then to test students whether they know the techniques to solve these bits. But today individuals create value by synthesizing the disparate bits. This is about curiosity, open-mindedness, making connections between ideas that previously seemed unrelated, which requires being familiar with and receptive to knowledge in fields other than our own. If we spend our whole life in a silo of a single discipline, we will not gain the imaginative skills to connect the dots where the next invention will come from.

Perhaps most importantly, in today's schools, students typically learn individually and at the end of the school year, we certify their individual achievements. But the more interdependent the world becomes, the more we rely on great collaborators and orchestrators who are able to join others in life, work and citizenship. Innovation, too, is now rarely the product of individuals working in isolation but an outcome of how we mobilize, share and link knowledge. Future tests should not disqualify students for collaborating with other test-takers, but encourage them to do so.

It is important to get this right. Tests influence policies and practices in many ways by signalling priorities for curriculum and instruction, and tests can focus the content of instruction when school administrators and teachers pay attention to what is tested and adapt curriculum and teaching accordingly.

Recent developments in assessment methodologies enable us to bridge the gap between summative and formative assessments, which has traditionally divided educators into two opposing camps. It is now possible to create coherent multi-layered real-time assessment systems that extend from students to classrooms to schools to regional to national and even international levels, and that provide dynamic task contexts in which prior actions stimulate unpredictable reactions that in turn influence subsequent strategies and options. Such tests can provide a window into students' understandings and the conceptual strategies a student uses to solve a problem, and they can add value for teaching and learning, when tasks incorporate transfer and authentic applications and provide opportunities for students to organize and deepen their understanding through explanation and use of multiple representations. Not least, such tests can feed results back to learners and educators in real time, so that data become a powerful instrument to improve student learning outcomes. Teachers can understand what the assessment reveals about students' thinking. And school administrators, policymakers and teachers can use this information to create better opportunities for student learning.

4

Bob Sornson

The way we test students and use that information reflects the systems architecture and the underlying design for how instruction will be delivered in our schools. For more than a century and a half we have used a systems model that was designed as a time-bound, age-based, one-size-fits-all curriculum-driven instructional system. We deliver instruction, test students, decide who excelled and who did not, offer grades, and move on to the next lesson. In answering the question, How should we test our students?, our curriculum-driven system responds primarily by using end-of-unit or standardized tests to efficiently sort students into winners and losers.

For essential skills and knowledge, solid competency is the goal, not an 80 per cent score on a one-time assessment.

During the mid-nineteenth century the Prussian, British, American, and other Western models of instruction were designed to 'expose' content to a large number of students, often for a relatively short period of months or years, and identify a small number of students who should be encouraged to continue in school and develop greater academic skills. A standard curriculum was used for each grade or course. There was little focus on trying to help all students stay in school or develop advanced academic skills. In an agrarian or early industrial society, a small number of highly educated people were sufficient for the management and academic needs of the society.

Testing was used to discern which students were more proficient in the material, identify good students from poor students, and encourage good students to continue their education.

In recent decades, academic skills have become far more important for access to economic success. But our schools are still designed to 'cover' standard material in a similar fashion for all students in a class or grade, 'test' students to determine which students were more proficient than others, and grade or 'sort' students into differing categories of success. So, as pressure to improve the learning outcomes of students increased, our 'Cover Test Sort' (CTS) education systems did more of what CTS systems are designed to do.

Schools began to 'cover' more content. States and nations began to formulate long lists of grade level and subject-area content expectations. Academic rigor somehow became defined as 'covering' more content in the time allowed. Teachers diligently attempted to 'cover' every chapter and deliver the entire prescribed curriculum. Rigid pacing guides and scripted learning programs were developed.

Schools began to 'test' with greater frequency and scrutiny. Mandated annual standardized assessment systems were devised to judge student success in keeping up with the standard curriculum. Using the same tests, we compared schoolwide results, compared communities and states, rated teachers, and declared more winners and losers. Many schools developed quarterly/monthly assessments to add additional pressure to ensure curriculum coverage, and a variety of other standardized assessments were added to the annual school schedule.

Most education systems now have an over-abundance of data to tell us which students, schools, and communities are winners and losers. But the limits of a CTS educational system are in its basic architecture. It was never built to help all children succeed. It was designed to cover a standard curriculum in approximately the same way for all students in an age group, give some tests, identify winners and losers, and move along to the next lesson.

The CTS system is well designed for identifying winners and losers, but not well designed to support the individual learning needs of each student.

In recent years, competency-based learning has emerged as an alternative systems model, offering personalized learning toward essential learning outcomes. Competency-based learning systems are built with a different framework for teaching, learning, and assessment. In a competency-based learning model, the systems architecture is designed to:

- understand each student's learning levels and learning needs;
- personalize instruction toward crucial skills or knowledge;
- maximize instructional gains by keeping students in their instructional zone;
- build solid skills that allow each student to continue learning for life.

To support progress toward crucial learning goals, careful formative pre-assessment of students gives teachers the information needed to know where to begin instruction, and if there are any important skills that must be addressed before grade-level instruction can begin.

Ongoing formative assessment is a crucial part of the competency model. Quizzes and tests may be part of this ongoing assessment, but teacher observation is always important for formative assessment. Teachers learn to embed opportunities for observational formative assessment into small group instruction, centers, projects, and other activities.

For essential skills and knowledge, solid competency is the goal, not an 80 per cent score on a one-time assessment. Proficiency must be observed or measured on several occasions, over a period of time, using several different learning contexts or materials before a teacher can certify that a student is truly able to understand and use this knowledge or skill.

Instead of using assessment to sort students, schools, teachers, communities, sub-groups and nations into winners and losers, in a competency-based system the primary purpose of assessment is to deeply understand each student so the teacher can design and deliver instruction that is well matched to his/her level of readiness. In a competency-based learning system the most important assessment is formative.

5

Dylan Wiliam

Testing plays an important role in raising achievement.

A test – or any other kind of assessment for that matter – is, at its heart, simply a procedure for making inferences. People engage in activities, we collect evidence about what they do, and on the basis of that evidence, we draw conclusions. So far, so good. However, because testing students takes time away from other things that we think are more important, we try to make tests serve multiple purposes. We use the same tests to find out what students have learned, what they need to do next, how good their teachers and schools are, and even how schools in one country compare with those in another.

As might be expected, this means that the tests don't serve any of the purposes very well, but – more seriously – the different purposes actually conflict. For example, national school examinations in England were originally developed by universities to inform decisions about which students to admit. However, over the last thirty years or so, examination results have increasingly been used as indicators of school quality, and specifically to inform parental choice.

Schools have therefore sought to increase their students' scores, often by 'spoon-feeding' their students. The result has been students getting higher and higher grades, but being less and less well prepared for advanced study.

Such 'gaming' of the system is possible because assessments are *samples* of the things we are interested in. Out of all the possible things we might ask students to do, we choose particular ones. Where the choices are predictable, performance on the assessment can be increased without a commensurate increase in the performance on the things that were not assessed. The result is learning that is *narrow, shallow* and *transient*. The learning is *narrow* because achievement on the tests can be increased by focusing on:

- some subjects rather than others (reading and mathematics are greater priorities than science or social studies, let alone art, music, dance or drama);
- some aspects of subjects rather than others (reading and writing is given more emphasis than speaking and listening; computation is given greater emphasis than using and applying mathematics);
- some students at the expense of others (greater attention is focused on students who are close to key benchmarks).

The learning is *shallow* because the kinds of things that can be tested in standardized conditions are limited. The mathematics curriculum might call for students to learn how to design, administer and analyse the results of a survey, but only trivial aspects of these processes can be assessed in a test. Finally, the learning is transient because what counts is a student's performance on the test itself, rather than six months later, and so there is little incentive to teach for long-term retention.

As students, teachers and schools have come under more and more pressure to increase students' scores on tests and examinations, it is perhaps not surprising that testing has come under criticism, with complaints that students are tested too much, and with demands that the amount of testing be reduced. This is unfortunate, because testing plays an important role in raising achievement. The best evidence we have suggests that student achievement is higher in education systems where students routinely take end-of-course tests and examinations.[ii] Perhaps more importantly, regular, frequent practice testing is one of the most cost-effective ways of increasing student achievement.[iii] The challenge, then, is to realize the benefits of testing while avoiding the significant negative effects that testing can have. The key idea here is that, though it is tempting to use the same information for multiple purposes, we must resist this temptation. Frequent regular practice testing benefits students, but students do not gain any additional benefit from a test when a score is recorded in a teachers' mark book. The best person to mark a test is the person who just took it. Similarly, electronic voting systems allow teachers to record every single student response to a class question, but if we want to create classrooms where students feel okay about making mistakes, the last thing we should do is record every single one of them.

Over the next decade or two, technology will change assessment in schools beyond recognition. Computers are already able to assess student essays more accurately than humans – an impressive sounding achievement until you realize how inaccurate humans are at this – but within a decade computers are likely to be able to build up sophisticated models of what students can and cannot do. Such models will be invaluable in helping teachers plan learning activities for their students, and it will be tempting to use such models to produce summary judgements of students' capabilities. I think this temptation must be resisted. If schools are to be

places where students experiment, take risks, learn from mistakes, then we should not seek to capture all the evidence. Assessment for summative purposes should be periodic, at the end of sequences of learning and designed to provide snapshots of a student's capabilities. At all other times, the focus must be on collecting evidence that will help students and teachers guide learning more effectively.

Notes

[i] Peterson, A. (1972) *The International Baccalaureate: An Experiment in International Education*, London, George Harrap,

[ii] Wiliam, D. (2010) 'Standardized testing and school accountability', *Educational Psychologist*, 45(2), 107–122.

[iii] Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J. and Willingham, D. T. (2013) 'Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology', *Psychological Science in the Public Interest*, 14(1), 4–58. doi:10.1177/1529100612453266.