Research Conference 2015

Learning Assessments
Designing the future

16–18 August 2015
Crown, Southbank, Melbourne

Australian Council for Educational Research
Contents

Foreword v

Pre-conference presentations Sunday 16 August

Creating stealth assessments 3
Professor Val Shute and Dr Michael Timms

Assessing young children’s literacy and mathematics understandings 5
Associate Professor Joanne Mulligan; Marion Meiers, Maurice Walker and Prue Anderson

Improving assessment for Indigenous students 8
Alison Quin

New measures for an old friend: A learning progression for ICT literacy 9
Professor Mark Wilson

Conference papers Monday 17 August

Keynote: Learning assessments: Designing the future 13
Professor Geoff Masters AO

Reflecting on teacher research on assessment: Challenges and innovations for the future 17
Professor Marie Brennan

Should generic curriculum capabilities be assessed? 18
Dr Rosemary Hipkins

Collaborative problem-solving: Assessment and reporting 23
Professor Patrick Griffin and Associate Professor Esther Care

PISA: Behind the headlines and past the rankings 24
Dr Sue Thomson and Chris Wardlaw PSM

Keynote: Assessment to action: New thinking from India 34
Dr Rukmini Banerji

Assessment standards, ‘intentional alignment’, and dialogic inquiry 40
Professor Claire Wyatt-Smith

Assessment in interactive learning environments 45
Dr Michael Timms and Dr Jason Lodge

Measuring what matters: Challenges and opportunities in assessing science proficiency 54
Professor James W Pellegrino
Conference papers Tuesday 18 August

Keynote: Stealth assessment in video games  
*Professor Val Shute*  
61

Hacking assessment  
*Associate Professor Phillip Dawson*  
65

The move to NAPLAN online: The advantages and the road ahead  
*Dr Stanley Rabinowitz with discussant Professor Barry McGaw*  
68

Translating rich learning assessments into certified results and university selection devices  
*Dr Gabrielle Matters*  
70

Assessing general capabilities  
*Julian Fraillon and Juliette Mendelovits*  
78

Poster presentations  
83

Conference program  
87

Crown floor plan  
93
Foreword
There is only one fundamental purpose of assessment in education. That purpose is to establish and understand where learners are in an aspect of their learning at the time of assessment.

— Professor Geoff Masters
It is indeed timely that Research Conference 2015 addresses the theme Learning Assessments: Designing the future. It is six years since our Research Conference considered issues in assessment, and the landscape is being significantly transformed. Not only is Australia’s school curriculum changing, but related issues of teaching quality and assessment practice are hot topics here and in many other countries.

This transforming landscape includes changes in thinking about the fundamental purposes of assessment; growing demands for the assessment of a broader range of student skills and capabilities; and new technologies that allow us to gather and visualise information about student learning more efficiently and thoroughly than ever before.

Whether we are teachers, researchers, leaders of schools or systems, we must not forget that improving learning is at the heart of assessment. As Dr Rukmini Banerji — a keynote speaker at this conference — says, assessment must be followed by action. Papers at Research Conference 2015 indicate that as we understand more about learning and pursue solutions to the issues we face, new challenges are emerging.

ACER has listened to many educators through our Rolling Summit on Assessment Reform and Innovation. We recognise that schools and nations are looking for sound evidence to inform their actions, and trust that this conference will provide you with both information and inspiration to contribute to designing the future of assessment.
Pre-conference presentations
Sunday 16 August
‘Stealth assessment is intended to be invisible and ongoing, to support real-time, just-in-time learning and to seriously reduce test anxiety, all without sacrificing assessment validity and consistency.’
— Professor Val Shute
Creating stealth assessments

Val Shute
Florida State University, USA

Val Shute is the Mack and Effie Campbell Tyner Endowed Professor in Education in the Department of Educational Psychology and Learning Systems at Florida State University. Before coming to FSU in 2007, she was a principal research scientist at Educational Testing Service, where she was involved with basic and applied research projects related to assessment, cognitive diagnosis, and learning from advanced instructional systems. Her general research interests hover around the design, development and evaluation of advanced systems to support learning, particularly related to 21st-century competencies. Her current research involves using games with stealth assessment to support learning of cognitive and non-cognitive knowledge, skills and dispositions. Her research has resulted in numerous grants, journal articles, books, chapters in edited books, a patent, and a couple of recent books, including Measuring and Supporting Learning in Games: Stealth Assessment (Shute & Ventura, The MIT Press, 2013) and Innovative assessment for the 21st century: Supporting educational needs (Shute & Becker, Springer-Verlag, 2010).

Michael Timms
Australian Council for Educational Research

Dr Michael Timms directs the Assessment and Psychometric Research Division at ACER, which develops high-quality assessments and conducts cutting-edge research in educational measurement. He is Chief Investigator in the Science of Learning Research Centre, in which ACER is a lead institution. He is a recognised leader in the development of innovative ways to assess students in electronic learning environments. His research is widely published in peer-reviewed journals and he was awarded the 2013 Journal for Research on Science Teaching Award by the National Association for Research on Science Teaching.

Prior to joining ACER, Dr Timms was Associate Director of the Science, Technology, Engineering and Mathematics (STEM) Program at WestEd, a pre-eminent educational research and development organisation in the United States. He led large-scale research studies in STEM education, with special focus on computer-based assessment projects, especially through the SimScientists research program (www.simscientists.org). He has been involved in the development of two assessment frameworks for the US National Assessment of Educational Progress for which he received the Paul Hood award for excellence in educational research at WestEd.

Dr Timms has experience in leading evaluation research projects for other educational research grant recipients, such as universities, and has managed large-scale item development projects across many content areas. He is knowledgeable about the education systems of Australia, the United States and the United Kingdom.

Dr Timms is an Associate Editor for the Australian Journal of Education.
Workshop

In this workshop, participants will learn how to create stealth assessments to measure student performance during interactions within computer-based learning environments, like digital games or intelligent tutoring systems. These measures are then used to estimate various competencies, including hard-to-measure constructs like creativity, persistence, problem-solving and systems thinking. First, we will explain how evidence-centred design can be used as a theoretical approach to designing such assessments. Next, we’ll illustrate how evidence-centred design was applied in the development of stealth assessment within a particular game (using the example of Plants vs. Zombies 2). Participants will have a chance to create their own evidence-centred design models (exploring competency, evidence and task), which can serve as an outline for an assessment related to any construct of interest. We’ll show how assessment of learning is implemented in the system using a particular method, Bayesian networks, or Bayes nets. Participants will learn how Bayes nets have been used to assess and support learning in different learning environments.
Assessing young children’s literacy and mathematics understandings

Collette Tayler
The University of Melbourne

Professor Collette Tayler will act as moderator for this symposium.

Professor Collette Tayler has held the Chair in Early Childhood Education and Care in the Melbourne Graduate School of Education since 2007. Professor Tayler is an educationalist; she conducts local and cross-national studies of the ways that social, family and educational policies and practices affect early childhood education and care outcomes. Her work addresses program access and engagement; public and private investments; program standards and quality; the curriculum and pedagogy applied in different services; leadership and staff development; child and family involvement and program outcomes. Her research seeks to explain both universal principles and contextual variations needed to provide exemplary care and education for young children. Professor Tayler holds a PhD in education from the University of Western Australia.
Abstract

Current research indicates that young children are capable of developing mathematical concepts and reasoning much earlier than previously considered. The development of mathematical concepts emerges out of children’s understandings about pattern and structure from interaction with the real world. An early mathematical assessment interview, the Pattern and Structure Assessment (PASA) focuses on a range of concepts and processes and is linked with mathematical attainment in the ACER Progressive Achievement Tests in Maths (PATMaths).

ACER is piloting early-years technology-based tools, such as the Digital Early Reading and Mathematics Assessment (DERMA). DERMA is an audio-based assessment of early reading and mathematics skills delivered on offline tablets fitted with headphones. The practice program enables young students who have never used a computer before to gain sufficient skill to independently work through the assessment. DERMA promotes the message that reading is a meaningful activity with tasks that model good classroom practices. It also promotes authentic mathematics tasks where students can drag and drop objects to count, sort and compare groups and order numbers. DERMA has been successfully used in Afghanistan in 2014 in a large-scale survey conducted in Dari and Pashtu with Grade 3 students. It has also been successfully used in pilot studies in Lesotho and with remote Indigenous communities in Australia. ACER is currently piloting a version of DERMA for use with Australian students in the first three years of school. This workshop session will demonstrate DERMA and cover the five main reasons for using DERMA: motivation of the students, portability of the medium, validity and content coverage, reliability, and efficiency of gathering sample data.
Abstract

The ACER 1999–2005 Longitudinal Literacy and Numeracy Study (LLANS), and several subsequent studies, investigated growth in literacy learning in the early years of school. A key finding was the wide distribution of achievement at school entry and the next two years at school. A further study with a random sample of Australian children drawn from preschools in the year prior to school entry, the LLANS Transition from Preschool to School, has investigated literacy development from preschool to Year 1. Following the LLANS methodology, one-to-one interview assessments were based on tasks and activities designed to provide information about students’ achievement in critical aspects of literacy: comprehension, reading fluency, phonemic awareness and phonics, concepts about print, and writing. Data were gathered from the preschool students in mid-2012, at the beginning of Foundation in 2013, and in Year 1 in 2014. A wide distribution of achievement across these three years was evident. For example, in preschool, 49 per cent of children knew where to start reading a story, rising to 77 per cent among students in Foundation. These assessments have clear potential for identifying starting points for teaching and planning to support progress in learning.
Alison Quin, who descends from the Tagalak people of the Gulf country in north western Queensland, has worked in education her entire career, obtaining a Graduate Diploma in Secondary Education and working in a variety of Indigenous community education fields. Alison was an Associate Lecturer in Indigenous Education for several years at the Centre for Indigenous Studies at Charles Sturt University, teaching Education students how to develop culturally inclusive practice, including the development of units of work and appropriate assessment. Recently, Alison has worked as a Learning Advisor, helping Indigenous university students develop their academic skills and succeed in assessment tasks, and has now moved into the field of Learning Design for tertiary courses at Queensland University of Technology. Alison’s passion is developing Indigenous culturally inclusive online learning contexts.

Workshop

This workshop will introduce a teaching scenario and existing unit of work built around Indigenous culturally inclusive practices. Participants will work in small groups to devise assessment items in a scaffolded process. The development of assessment tasks will be informed by clarifying Indigenous education, the process of building relationships between schools and Indigenous communities that contribute to co-developed units of work, and principles of Indigenous culturally inclusive practice such as place-based, community-grounded, group production.
New measures for an old friend: A learning progression for ICT literacy

Mark Wilson
The University of California, Berkeley and the University of Melbourne

Mark Wilson is Professor of Education at the University of California, Berkeley, and also at the University of Melbourne. He received his PhD degree from the University of Chicago in 1984. His interests focus on measurement and applied statistics, and he has published more than 100 refereed articles in those areas. Recently he was elected president of the Psychometric Society, and is currently Vice-President of the United States National Council for Measurement in Education (NCME); he is also a Member of the US National Academy of Education, a Fellow of the American Educational Research Association, and a National Associate of the US National Research Council. He is Director of the Berkeley Evaluation and Assessment Research (BEAR) Center. His research interests focus on the development and application of sound approaches for measurement in education and the social sciences, the development of statistical models suitable for measurement contexts, the creation of instruments to measure new constructs, and scholarship on the philosophy of measurement.

Professor Wilson is a member of the editorial board of the Australian Journal of Education.

Workshop

This workshop will present new thinking and new results from the work of the Assessment and Teaching of 21st-Century Skills (ATC21S) project, on the learning progression for information and communications technology literacy — learning in digital networks. This project, initially sponsored by Cisco, Intel and Microsoft, aimed to help educators around the world enable students with the skills to succeed in future career and college goals. The workshop will be structured to show how the development of the new ideas and measures for ICT literacy followed the logic of the assessment system developed by the Berkeley Evaluation and Assessment Research (BEAR) Center. The initial concepts behind the new measures are based on a recounting of the multiple changes in the conceptions of ICT literacy over the last 30 years, leading to the development of the new ICT literacy learning progression. This is followed by a discussion of the development of a set of interactive and group tasks that tap into the dimensions and levels of the learning progression, in the context of a web-based environment. A brief demonstration of two of the tasks will be a part of the workshop. Data were collected in this digital environment in four countries: Australia, Finland, Singapore and the United States, and these data will be used to explore the empirical underpinnings of the tasks and the learning progression. Ample opportunity for questions and discussion will be provided.
Conference papers
Monday 17 August
‘Children’s learning is an issue of national importance and therefore all assessment must lead to concrete action.’
— Dr Rukmini Banerji
Learning assessments: Designing the future

Abstract

Processes for assessing student learning are undergoing fundamental transformation.

This presentation will consider three developments which can be expected to shape how student learning is assessed in the future. First is fundamental change in how assessment is conceptualised and approached, with a focus on monitoring learning. Second is growing interest in the assessment of a broader range of skills and attributes than those addressed in most current assessment efforts. Third is advances in technology which are opening the door to new ways of gathering information about student learning, including through records of real-time interactions in online learning environments. In ACER’s Centre for Assessment Reform and Innovation, these three developments are referred to as ‘new thinking’, ‘new metrics’ and ‘new technologies’. This presentation will explore ways in which these three developments, together with scientific advances in our understanding of learning itself, can be expected to transform school assessment processes over the next decade.
Processes for assessing student learning are undergoing fundamental transformation.

Michael Barber and Peter Hill in their recent paper write of a coming ‘renaissance’ in educational assessment. Many of the forces for change that Barber and Hill identify are also described in my 2013 paper ‘Reforming Educational Assessment: Imperatives, Principles and Challenges’.

So what are these forces for change, and how will they shape the future of assessment?

Three developments underpin the transformation now underway. First, fundamental changes are occurring in how we conceptualise and approach the assessment of student learning. Second, there is growing international interest in, and demand for, the assessment of a broader range of skills and attributes than those addressed in most current assessment efforts. Third, advances in technology are opening the door to new ways of gathering information about student learning, including through records of real-time interactions in online learning environments.

In ACER’s Centre for Assessment Reform and Innovation we refer to these three developments as new thinking, new metrics and new technologies of assessment.

**New thinking**

In the past, assessment results in education have been used largely to judge and grade. This use of assessment is consistent with the view that the role of teachers is to teach, the role of students is to learn, and the role of assessment is to establish how well students have learnt what they have been taught — and to grade them accordingly. When used in this way, learning assessments are often viewed as straightforward and unproblematic.

It has become common to refer to the multiple ‘purposes’ of assessment. But in designing learning assessments for the future, a conceptual breakthrough is made by recognising that there is only one fundamental purpose of assessment in education. That purpose is to establish and understand where learners are in an aspect of their learning at the time of assessment.

The question of where learners are in their learning can be addressed for individuals and also for groups. It can be addressed at different levels of precision and in varying degrees of diagnostic detail. It is an essential question for classroom teachers, but is also crucial for education policymakers and system managers.

Information about where students are in their learning is necessary for identifying appropriate starting points for action. Teachers require information about starting points to target teaching on individuals’ levels of readiness and learning needs and to set appropriate stretch goals for further learning. But decision-makers at all levels — from students and parents to school leaders to system managers and governments — require dependable information about current levels of achievement to guide future action.

Information about where students are in their learning also is essential for monitoring learning progress over time. Success in learning is best defined and measured as the progress (or growth) that students make. Information about progress is required to evaluate the effectiveness of teaching strategies, but is equally crucial for evaluating initiatives to raise national achievement levels and close achievement gaps.

Under this way of thinking, the focus of assessment is on understanding the current situation and then using this understanding to guide future action, monitor progress, and evaluate the effectiveness of interventions. It has much in common with the use of assessment in other professions such as medicine and psychology, where the purpose is not so much to judge as to understand.

**New metrics**

Around the world, school curricula are giving greater emphasis to skills and attributes believed to be important for life and work in modern society. These skills and attributes — sometimes referred to as general capabilities, cross-curricular skills or 21st-century skills — include literacy and numeracy, problem-solving, oral communication, critical and creative thinking, the ability to work in teams, self-management and intercultural understanding. The growing use of new digital tools is requiring new capabilities in information and communications technologies, including new skills in reading, communicating, online searching and problem-solving.

Greater priority also is being given to students’ deep understandings of school subjects and their ability to apply those understandings to practical, real-world problems. This is sometimes referred to as a ‘literacy’ perspective. For example, ‘scientific literacy’ is defined as the ability to apply scientific knowledge and an understanding of scientific concepts and principles to everyday situations and problems.

These developments introduce a number of assessment challenges. First, considerable work is required to clarify newly-prioritised aspects of learning such as creative thinking and collaborative problem-solving. Can skills and attributes of these kinds be treated as general competencies or do they have meaning only in the context of specific school subjects? Related questions arise about the focus of assessment. For example, should an assessment of ‘teamwork’ focus on how well an individual works in and contributes to a team, or focus on the work of the entire team?
Second, constructs of these kinds usually require assessment methods very different from those used to assess mastery of curriculum content. Many require direct observations of learners’ performances in complex situations, possibly working collaboratively to solve real problems, to apply what they have learnt, think critically, create new solutions, communicate with others, and make effective use of technology.

Third, because general capabilities such as critical thinking, self-management and intercultural understanding develop throughout the years of school, assessment processes must be capable of monitoring students’ long-term development. The same is true of deep understandings of concepts and principles, which often develop only over extended periods of time. The implications are that assessment processes must be underpinned by pictures of what long-term improvements in these skills, attributes and understandings look like — that is, by learning ‘metrics’ for monitoring progress across multiple years of school.

New technologies

Advances in technology have the potential to transform assessment practice through more personalised, more interactive and more intelligent forms of evidence gathering, as well as by providing more immediate, high-quality feedback to learning processes.

Technology is providing enhanced learning and assessment environments. For example, in school science classes, students are manipulating variables such as forces, angles, distances and time and observing the effects of these changes in virtual environments that are sometimes difficult or impossible to create in normal classrooms. They are conducting on-screen experiments and recording and analysing their observations and measurements electronically. Such technologically enhanced environments provide unique opportunities to collect evidence about students’ knowledge and understandings, including by tracking the processes they follow in attempting to solve problems.

Technology also is enabling more personalised forms of assessment in which tasks are matched automatically to the real-time performances of individual students. By selecting tasks at an appropriate level of difficulty, ‘computer-adaptive’ assessments of this kind provide more relevant assessment experiences and superior information about where individuals are in their learning. Students also can be given greater control over assessment processes; for example, by choosing where and when they wish to be assessed.

Finally, it is possible to build into digital assessments expert knowledge about common student errors and misunderstandings and to use this knowledge to automate diagnosis and guidance. For example, if a student when prompted to add the fractions $\frac{2}{3}$ and $\frac{1}{4}$, gives the answer $\frac{3}{7}$, an automatic hypothesis could be generated about the process the student followed. This hypothesis could be tested by assigning other fractions addition tasks. The results of such exploration could be flagged for the teacher’s attention and/or lead to electronic tutoring in adding fractions.

A professional challenge

These developments, together with scientific advances in our understanding of learning itself, can be expected to transform school assessment processes over the next decade.

References


Dylan Wiliam
Institute of Education, UK

Dylan Wiliam is Emeritus Professor of Educational Assessment at the Institute of Education, University of London where, from 2006 to 2010 he was its Deputy Director. In a varied career, he has taught in urban public schools, directed a large-scale testing program, served a number of roles in university administration, including Dean of a School of Education, and pursued a research program focused on supporting teachers to develop their use of assessment in support of learning.
Reflecting on teacher research on assessment: Challenges and innovations for the future

Marie Brennan
Victoria University

Professor Marie Brennan is currently Professor of Education in the College of Education at Victoria University, Melbourne. She started her education career as a technical teacher of humanities in the 1970s, and has worked in many positions since then, including as a member of the Access Skills Project Team in the Curriculum and Research Branch, co-coordinating the Victorian School Improvement Plan, and as a policy analyst in the ministry-wide Policy Coordination Division. Her university positions include stints at Deakin University, Central Queensland University, the University of Canberra and the University of South Australia. Now back in Melbourne, Professor Brennan is active in research that involves teachers, students and community members, including in Australian Research Council projects in Queensland and South Australia regional areas, as well as in Melbourne’s west. She was the rapporteur at ACER’s recent Excellence in Professional Practice Conference (EPPC), where teachers presented reports of their improvement projects relating to assessment.

Abstract

The theme of the 2015 Excellence in Professional Practice Conference (EPPC), held by ACER in May, was Improving assessments of student learning. A review of content and presentations for EPPC showed up some important trends and issues for further analysis and discussion.

In a significant number of schools, the focus on assessment was well integrated into an overall school improvement plan or approach. In other schools, the school process overwhelmed the focus on assessment, and at times even overshadowed the focus of the study to be shared. One group of presentations focused on using existing instruments in ways that advanced the teachers’ understanding of the issue and students’ knowledge and learning needs. Another group of studies demonstrated creative, practice-driven teacher research, largely by individual teachers, that had resulted in key advances in assessment practice.

Renowned education thinker Lawrence Stenhouse in 1975 defined research as ‘systematic inquiry made public’. This presentation will argue that the advances documented at EPPC meet this definition, by demonstrating knowledge production that is not merely for self-consumption.

The presentation will suggest key issues for supporting teacher research, and explore key issues still in need of research and innovation in the field. There is a critical place for teacher research among the range of assessment, pedagogy and curriculum integrations which can be developed through a range of research methodologies. Practices for classrooms may well be dreamt up outside those classrooms — yet unless and until those practices are co-produced by teachers and students, they remain in a black hole, unpractised. There is thus a central and creative role for teachers in developing assessment practice, whether or not the innovations are supported by research. Indeed, the best research useful for teachers and their students around assessment may well identify the in-practice problems which are yet to be creatively addressed.
Should generic curriculum capabilities be assessed?

Dr Rosemary Hipkins is Chief Researcher at the New Zealand Council for Educational Research. She began her career as a secondary science and biology teacher and worked in teacher education before moving to NZCER. Rose was actively involved in the development of New Zealand’s current national curriculum framework and has led national research projects related to both curriculum and assessment innovation in New Zealand. She is interested in deepening understandings of the key competencies and has co-led the development of resources to support their meaningful implementation across the curriculum.

Abstract

Both Australia and New Zealand have recently taken up the idea of ‘key competencies’ (‘capabilities’ in the Australian national curriculum) initially proposed by the Organisation for Economic Co-operation and Development. In both countries we have made them our own by adapting them to suit our own educational contexts. People often say that these capabilities won’t be taken seriously unless they are assessed. So whether, and how, to assess them continue to be vexed questions. In this paper I argue that capabilities are more appropriately seen as changing the curriculum rather than adding to it. If we are serious about preparing students for the future, outcomes for learning need to be re-imagined at the complex intersection of capabilities and traditional content prior to determining any assessment approaches.
Key points

- Capabilities can be used as ‘ideas for teachers to think with’ as we re-imagine a curriculum for the future.
- Commentary on a curriculum for the future places increased emphasis on the quality of intellectual activity and on being able to use new learning in authentic demonstrations of capability (that is, real tasks where students choose and justify the best course of action, actively employing their new knowledge and skills).
- Assessment challenges include: providing opportunities for metacognition (students demonstrate their awareness of competencies in use); managing evidence derived in group contexts (learning is distributed); and aggregating multiple instances of competency demonstrations (opportunities vary and different aspects of each key competency are called into play in different contexts).
- Annotated e-portfolios provide one practical means of addressing all these challenges, but their effective use is reliant on the development of rich tasks that allow students to demonstrate their growing competency levels.

Introduction

Should capabilities be assessed and if so how? This has been a vexed question since the inception of so-called ‘21st-century’ national curricula in both Australia and New Zealand (Hipkins, 2007). These capabilities were introduced as one part of a curriculum framework intended to bring teaching and learning into the present, so there is an important prior question about the curriculum ‘work’ they are expected to do. In our most recent research we have found it useful to encourage school leaders and teachers to think about key competencies as ‘ideas to think with’ (Hipkins, Bolstad, Boyd & McDowall, 2014). If we restrict our thinking about capabilities to ‘things students should get more of’ it is too easy to fall back into familiar outdated curriculum assumptions and miss the profound change potential in the very idea of building capabilities.

Rethinking purposes for learning

In previous curricula, the acquisition of knowledge and skills was largely taken as a given for assessment programs and practices. However, rapid social and economic changes, along with ever-more rapid evolution of uses and demands of digital technologies, have greatly expanded the range of types of outcomes learners need to achieve to be active participants in modern life. A recent analysis of ‘21st-century’ competency-based frameworks identified four common sets of outcomes: collaboration; communication; literacy in information and communication technologies; and social and/or cultural skills and citizenship. Most frameworks also mentioned: creativity; critical thinking; problem-solving; and development of quality products/productivity (Voogt & Pareja Robin, 2012).

The Australian Curriculum capabilities and the New Zealand Curriculum key competencies point towards these sorts of outcomes. The challenge is that they do not indicate how these types of outcomes are related to the knowledge and skills of the traditional curriculum. They might still be seen as adding to the curriculum — something to be assessed on top of (or instead of) traditional content. This understanding has led many schools in New Zealand to develop over-simplified rubrics for assessing key competencies as if they were generic personality traits of individual students. In my view this is neither appropriate nor fair, for reasons I will outline in the rest of the paper (Hipkins, 2009).

Developing reciprocal relationships between capabilities and traditional subject-based learning

Thinking differently about the relationship between capabilities and traditional subject-based curriculum content is helpful, but is not necessarily easy to do. We recently developed a suite of ‘engaging examples of practice’ that illustrate ways to integrate the New Zealand Curriculum key competencies into subject learning. Leading teachers were our inquiry partners in this applied research. All the examples the teachers helped us to shape demonstrate strong learning benefits when reciprocal relationships between the key competencies and more traditional subject area learning are strategically leveraged.

We noticed that all these teachers were thinking about two ‘layers’ of outcomes for the learning they designed. They had immediate goals (typically specific knowledge and skills) but they also had in mind longer-term goals — things they hoped students would become or be able to do in their futures (for examples see Hipkins & McDowall, 2013). The pedagogy they employed was critical to how they opened up opportunities for students to become more capable. This suggests that outcomes for learning need to be re-imagined prior to determining any assessment approaches. It also suggests that what teachers do to support capability development is as important as what students do. With this challenge in mind, we developed a self-audit framework to help teachers evaluate whether they were providing effective learning opportunities to support their students’ capability development.
How could demonstrations of capability be assessed?

The assessment challenge changes when learning opportunities are re-imagined, but it doesn’t go away. We still need a broad guide to the types of assessment tasks that could show the intended learning was successfully achieved. The following principles were distilled from multiple research–practice partnerships over the last decade, for a project that explored the question of whether and how we might assess students’ development of ‘international capabilities’ (Bolstad, Hipkins & Stevens, 2014). These principles offer a guide for thinking about assessment task design and the type of data that might be captured.

Principle 1: Assess competency in action

Re-imagining learning as a ‘complex performance’ (Hipkins, Boyd & Joyce, 2005) brings together the content, the context and the targeted capabilities to undertake a rich task. Note that all the capabilities will be woven into a coherent whole in any one task situation. It follows that whichever of the capabilities is least developed will likely limit what students are able to do. Rich tasks will often cross curriculum boundaries. This presents a greater challenge for designing learning experiences and assessments for secondary students than for primary students. Another challenge is that some aspects of capability are best enabled and demonstrated in group settings. Collaboration is an obvious example. Traditionally assessment judges the performance of an individual, regardless of how well the context enables or constrains that performance (in this case how well group dynamics allow collaboration to actually be demonstrated). Yet another challenge is that collaboration in modern contexts is often virtual rather than face-to-face. This brings its own complex demands to engage in interactions with others who have different perspectives, negotiate shared meanings, and co-construct problem resolutions, all within virtual spaces (Dede, 2009).

Principle 2: Collate evidence from multiple sources

Performances can be variable for a range of reasons. This creates issues of validity and reliability, as these are traditionally understood. An implication is that more than one source of evidence will be needed. In any case, one indicator of stronger capability is that what the student knows and can do can be adapted and transferred from simpler to more demanding contexts. Some rich opportunities to demonstrate capability will be available in settings beyond the school. But how to gather, moderate and add that evidence to a record of learning is a challenge that most schools have yet to address. Aspects of the chosen context for a performance can impact differently on different students’ abilities in demonstrating their capabilities — their backgrounds and prior learning experiences can help them see the action possibilities in a task, or not. This means it will be important to take identity, language and culture into account, both when designing assessments and interpreting their results.

In essence, we need to design systematic ways to record learning achievements from multiple sources, including different contexts, and to keep this record building over time. Possible approaches include development of annotated portfolios of evidence or learning logs. These allow an assessor’s observation of an authentic performance to be combined with a degree
of self-assessment. Learner input enables the assessor to include consideration of what the student was trying to achieve in the performance being judged (see also Principle 3).

**Principle 3: Involve students in assessment decision-making**

It is important to design assessment approaches that engage and involve students in gathering and reflecting on the evidence of their learning and growth. It’s often said that we need learning approaches for the future so that students become ‘lifelong learners’. Involving students in assessing their own learning makes a strong contribution to this future-focused aspiration. Each student needs to build their own ‘assessment capability’ (Absolum, Flockton, Hattie, Hipkins & Reid, 2009; Booth, Hill & Dixon, 2014). This enables students to get better and better at judging the quality of their own work, understanding assessment feedback, and seeing the big picture of what that feedback can and cannot tell them about their performance. The achievements of student athletes and their coaches show that this can be done — but both parties to the learning have to work at it.

Another reason to involve students in assessment is that developing metacognitive awareness of one’s current capabilities and next learning challenges is an important aspect of stretching and strengthening all the capabilities (Hipkins, 2006). It’s not enough to use current capabilities intuitively if we want to build adaptive expertise (that is, the ability to consciously change how we deploy our capabilities when the context or task require this).

If rubrics are used, students should be involved in conversations about their meaning, and take an active part in the judgement being made. Ideally, they would also be involved in constructing the rubrics in the first place. However, many questions still surround the nature of progression in capability development, so careful attention would need to be paid to any assumptions about the nature of progress being captured in the rubrics.

**Implications**

Experience in New Zealand schools suggests that it will be very demanding to design effective new curriculum and assessment tasks that encapsulate the principles outlined above. This will need to happen right across the curriculum, and at all levels of schooling. All teachers will need opportunities to take part in rich professional learning that unsettles tacit assumptions about purposes for learning and revisits the very idea of capabilities in a more expansive framing. The challenges for student learning and capability development apply to teacher learning too. Senior leaders need to be strong leaders of professional learning (Robinson, Hohepa & Lloyd, 2009) and make space for teachers to work collaboratively as they re-imagine a curriculum for the future.

**References**


Collaborative problem-solving: Assessment and reporting

Patrick Griffin
The University of Melbourne

Emeritus Professor Patrick Griffin held the Chair of Education (Assessment) and was Director of the Assessment Research Centre at the University of Melbourne from 1996 to 2015. He was also the Associate Dean in the Melbourne Graduate School of Education. He has published widely on assessment and evaluation topics that include competency development, language proficiency, industrial literacy, school literacy and numeracy, professional standards and online assessment and calibration. His recent publications include Assessment for Teaching, which details his work on developmental or growth models of assessment and learning (Cambridge University Press, 2014).

Professor Griffin was awarded the John Smythe medal for excellence in research for his work on profiling literacy development. He is a project team leader for UNESCO in southern Africa and was awarded, in 2005, a UNESCO Research Medal by the Ministers of Education from southern African nations. Professor Griffin is a World Bank consultant in Vietnam. He is a fellow of the International Academy of Education, the Australian College of Educators and the Australian Council for Educational Leadership. Over the past 40 years he has addressed major professional associations, taught, and conducted assessment and evaluation research projects in more than 20 countries.

He is currently working with the Victorian state education department and is leading several large studies funded by the Australian Research Council. He is a member of the Measurement Expert Advisory Group for the National Assessment Program — Literacy and Numeracy. He remains as the Executive Director of the Assessment and Teaching of 21st Century Skills project — a multi-year, multi-country public–private–academy partnership project originally sponsored by Cisco, Intel and Microsoft. The work of this project has been published in two volumes, for which Professor Griffin is the lead editor (Springer International Publishing, 2015).

This conference session will be co-presented by Associate Professor Esther Care, Deputy Director of the Assessment Research Centre in the Graduate School of Education at the University of Melbourne. Associate Professor Care is the International Research Coordinator for the Assessment and Teaching of 21st Century Skills project.

Abstract

This practical session will present a live administration of interactive collaborative problem-solving assessment and reporting. The presenters will demonstrate example tasks and reports. Audience representatives will have the opportunity to role-play as students being assessed across a range of social and cognitive skills associated with collaborative problem-solving. The discussion will then explore how these social and cognitive skills can be incorporated into the teaching program, enabling higher-order skills to be assessed in key learning areas across the curriculum.
Sue Thomson
Australian Council for Educational Research

Dr Sue Thomson is Director of the Educational Monitoring and Research Division at the Australian Council for Educational Research and Chief Investigator in the Science of Learning Research Centre, in which ACER is a lead institution. She is also Research Director for the National Surveys research program at ACER, overseeing Australia’s participation in all international and national sample surveys.

Dr Thomson has also fulfilled the roles of National Research Coordinator for Australia in the International Association for the Evaluation of Educational Achievement (IEA) Trends in International Mathematics and Science Study (TIMSS) since 2002, National Project Manager for Australia in the Organisation for Economic Co-operation and Development (OECD) Programme for International Student Assessment (PISA) since 2004, and National Research Coordinator for Australia in the IEA Progress in International Reading Literacy Study (PIRLS) since 2008.

Dr Thomson’s research at ACER has involved extensive analysis of large-scale national and international data sets — the Longitudinal Surveys of Australian Youth (LSAY), as well as TIMSS and PISA. She is involved in several projects involving analysis of the longitudinal data collection associated with the PISA surveys. She was engaged as an expert writer on the National Numeracy Review, and has consulted with a variety of government departments at both Commonwealth and state levels, as well as with the Catholic Education Commission, on a variety of data-analysis projects related to TIMSS and PISA. Dr Thomson is an Associate Editor for the Australian Journal of Education.

Chris Wardlaw PSM
Victorian Curriculum and Assessment Authority

Chris Wardlaw PSM is currently Chair of the Victorian Curriculum and Assessment Authority. Chris held a Deputy Secretary position in education in Hong Kong (2002 to 2008) and Victoria (2009 to 2013) and is now retired. In the Hong Kong Government Chris had responsibility for curriculum, assessment and quality assurance for pre-primary, basic education, senior secondary education and the interface with tertiary education, and in Victoria for strategy and review across the portfolio.

Chris has considerable experience authorising, funding, and learning from national and international assessments in a range of domains across jurisdictions. Hong Kong has improved its already high standing in successive international assessments from 2000 to the present.

Chris has had an extensive career in Victorian education during which he took a leading role in major reforms supporting school-level decision-making and evaluation and review.

Chris was awarded the Public Service Medal (PSM) in the 2013 Queen’s birthday Honours list and was made a fellow of Monash University in 2013.

In a parallel sporting career, Chris was Head Coach of the Australian Athletics Team at the 2000 Sydney Olympic Games; an Olympian in 1976 and 1980 in the 10 000m and marathon events; and coach of marathoners Steve Moneghetti and Kerryn McCann and distance runner Craig Mottram. Chris was awarded the Australian Sports Medal in 2000.
**Abstract**

Whenever the results of the Programme for International Student Assessment (PISA) are announced, media headlines are full of reports about rankings, about how many countries Australia is outperformed by and outperforms. In early rounds of PISA, Australia ranked among the top 10 countries across all three education domains assessed. However, over time Australia’s position has declined, rather than improved, and Australia no longer sits in the top 10 of any of the assessed domains.

This presentation will go behind the headlines and past the rankings, to look at where Australia has declined, and look at how we can improve outcomes for students and achieve a world-class education system.

In particular this presentation will focus on mathematics.
Whenever the results are released from one of the international assessments, the Programme for International Student Assessment (PISA) in particular, the headlines are full of reports about rankings, about how many countries Australia is outperformed by and outperforms. PISA is part of the National Assessment Program, acting as a component of the evaluation of the Melbourne Declaration on Educational Goals for Young Australians, which in the preamble explained:

> Australia has developed a high-quality, world-class schooling system, which performs strongly against other countries of the Organisation for Economic Co-operation and Development (OECD). In international benchmarking of educational outcomes for 15-year-olds in the 2006 OECD Programme for International Student Assessment, Australia ranked among the top 10 countries across all three education domains assessed. Over the next decade Australia should aspire to improve outcomes for all young Australians to become second to none amongst the world’s best school systems. (MCEETYA, 2008)

However, over the following seven years and two further cycles of PISA, Australia’s position has declined, rather than improved, and Australia no longer sits in the top 10 of any of the assessed domains.

In the most recent assessment, PISA 2012, compared only to those countries that took part in PISA 2003 (years in which mathematical literacy was the major focus of the assessment):

- four countries significantly outperformed Australia in both cycles
- six countries whose scores were not significantly different to Australia in 2003 outperformed Australia in 2012
- three countries whose performance was significantly lower than Australia in 2003 scored at the same level as Australia in 2012
- two countries whose performance was significantly lower than Australia in 2003 significantly outperformed Australia in 2012.

Typical of headlines in Australia after the most recent PISA study was this one that asked: Australia’s PISA slump is big news but what’s the real story? (Riddle, Lingard & Sellar, 2013)

What is the real story? This presentation will go behind the headlines and past the rankings, to look at where Australia has declined, and look at how we can achieve what the ministers hoped in 2008.

In particular this presentation will focus on mathematics. The Australian Council of Learned Academies recommends that Australia needs to grow its pool in the area of science, technology, engineering and mathematics (STEM), and expanding this talent pool requires increasing the participation of young women, a resource that is at the moment underutilised (Marginson, Tytler, Freeman & Roberts, 2013). The Year 10 students in particular that are assessed as part of PISA are at a crucial stage in their education — ready to make decisions about the subjects they choose to study in senior secondary school and what careers they may go into. A strong influence on their decision-making will be what they are confident and interested in.

### Mathematical literacy

In each cycle of PISA, three main areas are assessed: reading literacy, mathematical literacy and scientific literacy. In each cycle the assessment areas are rotated so that one domain is the major focus (the major domain), with a large amount of the assessment time being devoted to this domain compared to the other two domains (the minor domains). Mathematical literacy was the major domain in the second PISA assessment in 2003, and as this was the first year that was the case, comparisons are generally made back to this date. Mathematical literacy was also the major domain of the most recent PISA assessment, in 2012.

As the headlines indicated, Australia’s average score has declined, from 524 score points to 504 score points, as shown in Figure 1. In both cycles this score is significantly higher than the OECD average, however in PISA 2012 this was because the OECD average had also significantly declined (from 500 to 494 score points — perhaps due to the inclusion of some low-performing countries in the OECD in the 2012 cycle). While it appears that there was a decline from one cycle to the next in Australia, it was only the decline from 2003 to 2012 that reached statistical significance.

![Figure 1 PISA mathematical literacy 2003–2012, Australia](image-url)
To examine whether this decline was for students of all abilities, or whether it was concentrated amongst students at particular levels of ability, the distribution of achievement for each PISA cycle was examined. The distribution of each cycle was described by five percentiles (the 10th, 25th, 50th or the median, 75th and 90th) and their associated standard errors. A percentile is the value of a variable, the PISA mathematics scale score in this instance, below which a certain per cent of the population fall. For example, in 2012, the 90th percentile in mathematical literacy was 630, which means that 90 per cent of the population scored below 630 on the PISA mathematical literacy scale.

Figure 2 shows that rather than a single decline in scores at any one point of the distribution, the decline has occurred more gradually over time across the whole distribution. The smallest decline was at the 90th percentile, however it was still a statistically significant decline of 14 score points since 2003. The largest differences were seen in the middle of the distribution: at the 50th percentile the decline was 24 score points, a little more than two-thirds of a year of schooling.1

In addition to mean scores, PISA attaches meaning to the performance scale by providing a profile of what skills and knowledge students have achieved. The performance scale is divided into levels of difficulty, referred to as proficiency levels. In mathematical literacy there are six proficiency levels described, ranging from low (Level 1):

- Students can answer questions involving familiar contexts where all relevant information is present and the questions are clearly defined. They are able to identify information and carry out routine procedures according to direct instructions in explicit situations. They can perform actions that are almost always obvious and follow immediately from the given stimuli. (Thomson, De Bortoli & Buckley, 2013)

...to high (Level 6):

- Students can conceptualise, generalise and use information based on their investigations and modelling of complex problem situations, and can use their knowledge in relatively non-standard contexts. They can link different information sources and representations and move flexibly among them. Students at this level are capable of advanced mathematical thinking and reasoning. These students can apply this insight and understanding, along with a mastery of symbolic and formal mathematical operations and relationships, to develop new approaches and strategies for addressing novel situations. Students at this level can reflect on their actions, and can formulate and precisely communicate their actions and reflections regarding their findings, interpretations and arguments, and can explain why they were applied to the original situation. (Thomson, De Bortoli & Buckley, 2013)

Figure 3 shows the proportion of high-achieving and low-achieving students in each cycle of PISA. High achievers are those students who achieved at Proficiency Level 5 or Proficiency Level 6; low achievers are the proportion of students who failed to meet Proficiency Level 3. The proportion of high achievers in mathematical literacy dropped from 20 per cent in 2003 to 16 per cent in 2006 and then remained relatively stable in 2009 and 2012. Overall, though, the proportion of high achievers in 2012 was significantly lower than in 2003.

At the lower levels of achievement in PISA 2003, 33 per cent of Australian students failed to meet the minimum proficient standard. In the PISA 2012 assessment, this had risen to 42 per cent of students, a significant increase.

In summary, Australia’s position overall declined significantly in mathematical literacy from PISA 2003 to PISA 2012. This decline has been right across the distribution of achievement levels, from high to low. While this decline has been fairly consistent across the distribution, there was a substantially larger proportion of students in 2012 at the lower achievement levels, resulting in four in ten students not achieving our own minimal proficient standard.

Trends in mathematical literacy performance by gender

According to news coverage following the release of PISA results, ‘Australian girls’ performance in maths has fallen to the OECD average — dragging down Australia’s result.’ (News Limited)

So this is where the blame lies! Is this indeed the case, and is it the whole story?

Internationally and in Australia, a vast body of research has been conducted into gender differences in mathematics over several decades. Campaigns in Australia that encouraged female students to undertake mathematics, in particular, seemed to have been largely successful. In the Trends in International Mathematics and Science Study (TIMSS) in 1994/95, Australia was one of the six countries that had no gender differences in mathematics for Year 8 students, and also was one of the countries that had equivalent results by gender in advanced mathematics at Year 12 (Lokan, Ford & Greenwood, 1997). In PISA 2003 only a few score points separated males and females, a difference that did not reach statistical significance. Both scored significantly better than the OECD average.

---

1 It is possible to estimate the score point difference that is associated with one year of schooling. This difference can be estimated for Australia as there are a sizeable number of 15-year-olds who were enrolled in at least two different year levels in the PISA 2012 sample. Analyses of these data indicate that the difference between two year levels is, on average, 35 score points on the PISA mathematical literacy scale. This implies that one school year corresponds to an average of 35 score points in Australia.
Figure 2 Distribution of mathematics achievement, all students, Australia, PISA 2003–2012

Figure 3 Percentage of high and low achievers, all students, Australia, PISA 2003–2012

Figure 4 Mean PISA mathematical literacy scores, Australia, by gender

*indicates significant difference
Figure 5 Differences in mathematical literacy score for males and females between PISA 2003 and PISA 2012, Australia, by percentile

Figure 6 Score point difference between males and females, 2003 and 2012

Figure 7 Percentage of high and low achievers, by gender, Australia, PISA 2003–2012
Over the period 2003 to 2012, the average score for both males and females declined significantly — by 17 score points for males and 24 score points for females (Figure 4). In PISA 2012 in Australia, males achieved a mean score of 510 score points, which was significantly higher than the mean score of 498 score points for females. This difference of 12 score points equates to around one-third of a year of schooling, and the average score for female students has declined to such an extent that it is no longer significantly different to the OECD average.

Figure 5 shows the difference in mathematical literacy scores between PISA 2003 and PISA 2012, for males and females separately. What we can learn from this is that for females the largest decline was amongst lower-achieving students — more than 20 score points at the 10th, 25th and 50th percentiles, while for higher-achieving students (those at the 90th percentile) the decline was only six percentage points. For males the decline was more general — 11 percentage points at the 10th percentile peaking in the middle, with the ‘average’ student’s score declining by 22 score points, and then at the 90th percentile a decline of 13 score points over the nine years.

Figure 6 presents the differences between male and female mean scores at each percentile. Female students at the very lowest levels of achievement outperformed their male counterparts by five percentage points. At the 25th percentile, there was negligible score difference between the two groups. At the 90th percentile, the difference was some 15 score points in favour of male students.

In 2012, several differences can be noted. From Figure 5 we know that the performance of females declined more than that of males, and so perhaps it is not surprising that in 2012 males outscored females at both the 10th and 25th percentiles, and while there was little change around the middle of the distribution, females at the 90th percentile had decreased the lead of male students from 16 score points to nine score points.

These findings are also reflected in changes in the proportions of male and female students reaching various proficiency levels (Figure 7). From 2003 to 2012, the proportion of female students not achieving the Australian proficient standard (Proficiency Level 3) grew from 33 per cent to 43 per cent. At the same time the proportion of males not achieving this level increased from 33 per cent in 2003 to 40 per cent in 2012. While it is of concern that the proportion of females at the lower levels of achievement has increased so far in nine years, it is of more concern that the performance of both males and females has declined to such an extent.

At the same time, at the higher levels of achievement, the proportion of both male and female students at Proficiency Level 5 or Level 6 has declined by about the same amount — from 22 per cent to 17 per cent for males and from 18 per cent to 12 per cent for females.

In summary, the overall decline in Australia’s score in mathematical literacy is a reflection of a decline by both males and females over the last ten years; however this has been more marked in female students. While the average score for males remains significantly higher than the OECD average, the score for females slipped to a level where it is not significantly different to that mean. However, the data also reveal that much of the decline for females has been at the lower end of the achievement distribution, with the gender gap at the highest percentile actually decreasing between 2003 and 2012. For both males and females, there are a larger proportion of students failing to achieve the minimum benchmark of Proficiency Level 2, and fewer achieving the higher proficiency levels.

Students’ motivation and engagement can have a profound impact on their classroom performance in the short term and can affect the quality of their learning in the long term. A number of attitudes have been examined in both PISA 2003 and PISA 2012, allowing an investigation of whether these have changed across time.

**Attitudes and beliefs: The value of context**

Past the rankings, PISA provides contextual information about students’ beliefs and attitudes about mathematics. Are there attitudinal differences between males and females that might help explain the differences in their achievement levels? In PISA students are asked to rate their level of agreement to a range of contextual questions, usually on a Likert scale ranging from strongly agree to strongly disagree, where each scale is constructed to have a mean over the OECD of 0 and a standard deviation of 1. Positive or negative values do not necessarily mean that students responded positively or negatively to the underlying questions, rather that they responded more or less positively than students on average across the OECD.

A summary of the mean index score for each of these in 2003 and 2012 for males and females is shown in Table 1. Scores for females on each one of the attitudinal variables is significantly lower than the equivalent score for males, in both 2003 and 2012. In both years, female students showed lower levels of intrinsic motivation, self-efficacy and self-concept in mathematics than those of their male counterparts and lower than the average for all students across the OECD. While none of these has changed over time it is likely that they all contribute to the big picture, and should be addressed.
Intrinsic motivation

Students’ level of intrinsic motivation was measured in PISA as the amount of interest or enjoyment students felt in relation to mathematics. Females responded less positively than males on every item in this scale. For example, on the item ‘I am interested in the things I learn in maths’, 46 per cent of females agreed or strongly agreed, compared to 61 per cent of males and an average of 53 per cent across the OECD. On average, Australian females scored more negatively than the OECD average while males were more positive, as a whole.

Instrumental motivation

In addition to being motivated by how much they enjoy the subject, students will also be influenced to participate in mathematics if they perceive it to be useful for their future. This was measured in PISA by four statements comprising the instrumental motivation to learn mathematics scale. An example of this: ‘Mathematics is an important subject for me because I need it for what I want to study later on’ gained agreement from 80 per cent of males and 67 per cent of females. In this instance, the scores for males and females were both significantly higher than the OECD average, but the score for boys was substantially higher than that for females, indicating males felt much more that maths would be useful for them.

Self-concept

Self-concept and self-efficacy can be thought of as constructs that relate to students’ competency-related beliefs at different levels of generality; mathematics self-concept relates to how confident a student feels in mathematics in general, while mathematics self-efficacy has to do with how confident a student feels in relation to particular mathematics tasks. Self-concept was assessed in PISA with statements such as ‘I learn maths quickly’, with which 62 per cent of males and 46 per cent of females agree, compared to the OECD average of 52 per cent.

The index scores for self-concept show that the average self-concept in mathematics of Australian females was significantly more negative than both the male students and the OECD on average.

Self-efficacy

Self-concept and self-efficacy are both forms of competency beliefs; however, self-efficacy is more specific and asks how competent students anticipate they will be on a defined task. For example, students in PISA 2012 were asked how confident they would be doing a variety of tasks, including ‘calculating the petrol consumption rate of a car’. This item showed the most difference in confidence levels of males and females in Australia, with 41 per cent of females saying they were confident or very confident of being able to calculate this, compared to 66 per cent of males and 54 per cent of students on average across the OECD.

Maths anxiety

Maths anxiety (or the worry or tension felt when confronted with mathematical tasks) can have a negative impact on students’ ability to demonstrate their potential in a subject. In PISA 2003 and PISA 2012 anxiety was measured by asking students their level of agreement with five statements:

- I often worry that it will be difficult for me in mathematics classes.
- I get very tense when I have to do mathematics homework.
- I get very nervous doing mathematics problems.
- I feel helpless when doing mathematics problems.
- I worry that I will get poor grades in mathematics.

This was the only one of the attitudinal variables listed in Table 1 on which scores changed from PISA 2003 to PISA 2012, and showed a significant increase in maths anxiety for females, making the already significant difference in scores for males and females even larger (Figure 8).

Figure 9 shows the level of maths anxiety for students in each proficiency level for PISA 2012. The overall pattern of this relationship is as would be expected, with higher levels of anxiety at lower levels of achievement and lower levels of anxiety at higher levels of achievement. Notable is that the anxiety levels of female students are higher than those of male students at each proficiency level, including Proficiency Level 6, where there is a substantial difference (0.7 of a standard deviation) despite there being no significant difference in the scores of male and female students.

So yes, the headline at the beginning of this section was correct — girls’ performance has declined, although to say it is dragging down Australia’s results is exaggerated. However, girls are performing well overall, given their level of belief and confidence in themselves. If, however, Australians believe in improving the achievement levels of all students, including females, there needs to be work done in the area of changing perceptions and dealing with the underlying causes of maths anxiety.
Table 1 Mean scores on attitudinal variables, PISA 2003 and PISA 2012

<table>
<thead>
<tr>
<th></th>
<th>2003 Mean</th>
<th>2003 Se</th>
<th>2012 Mean</th>
<th>2012 Se</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic motivation</td>
<td>-.10</td>
<td>.02</td>
<td>.18</td>
<td>.02</td>
<td>*</td>
</tr>
<tr>
<td>Instrumental motivation</td>
<td>.11</td>
<td>.02</td>
<td>.34</td>
<td>.02</td>
<td>*</td>
</tr>
<tr>
<td>Self-concept in maths</td>
<td>-.08</td>
<td>.02</td>
<td>.29</td>
<td>.02</td>
<td>*</td>
</tr>
<tr>
<td>Self-efficacy</td>
<td>-.09</td>
<td>.02</td>
<td>.28</td>
<td>.03</td>
<td>*</td>
</tr>
<tr>
<td>Maths anxiety</td>
<td>.09</td>
<td>.02</td>
<td>-.19</td>
<td>.02</td>
<td>*</td>
</tr>
</tbody>
</table>

Figure 8 Maths anxiety, PISA 2003 and PISA 2012, by gender

Figure 9 Maths anxiety by proficiency level, PISA 2012, by gender
Conclusions

Gender differences in mathematics are important. There is evidence from PISA that scores for females are declining at a faster rate than scores for males, and that on all attitudinal variables, female students have more negative perceptions than male students about their ability and capacity to do mathematics, and lower levels of enjoyment coupled with higher levels of anxiety, even when there is no possible reason to exhibit these characteristics. In purely economic terms, Hanushek and Woessman (2015) have calculated that if all students in Australia were to achieve the minimum OECD proficiency, Proficiency Level 1, there would be a possible 16 per cent increase in gross domestic product (GDP). One can only imagine the impact on the economy if all students were to achieve Proficiency Level 2.

PISA also highlights gender differences in reading scores (whereby females are outperforming males), and it is true that these differences are of a much larger scale than the gender differences in mathematics. However the most recent findings of the Programme for the International Assessment of Adult Competencies (PIAAC) provide the outcomes for adults in Australia in both reading and mathematics. The proficiency levels for each are shown in Figure 10, and show clearly that while males have well and truly caught up with females by adulthood in reading, this is not the case for mathematics. Female students in PISA are showing an enormous lack of engagement, and this is translating into them dropping mathematics as soon as they are able to, resulting in a continued decline in scores into adulthood.

The data derived from PISA are invaluable in terms of being able to see the big picture of how a system is faring against other systems internationally, systems whose students will enter the workforce and work in industries competing against those for which Australian students will eventually work.

The data are also invaluable within a country to see how different equity groups or subgroups within the population are faring over time. The data available are rich, and provide more than just the means and rankings, enabling educators and policy-makers to look more deeply into differences that are apparent.

References


Figure 10 Proficiency levels in reading and mathematics, PIAAC 2012

<table>
<thead>
<tr>
<th></th>
<th>Reading literacy</th>
<th>Mathematics literacy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Level 1</td>
<td>Level 2</td>
</tr>
<tr>
<td>Males</td>
<td>14</td>
<td>30</td>
</tr>
<tr>
<td>Females</td>
<td>14</td>
<td>31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 5/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>19</td>
<td>31</td>
<td>34</td>
<td>17</td>
</tr>
<tr>
<td>Females</td>
<td>25</td>
<td>35</td>
<td>31</td>
<td>9</td>
</tr>
</tbody>
</table>
Assessment to action: New thinking from India

Rukmini Banerji
Pratham and the ASER Centre, India

For the past 15 years, Dr Rukmini Banerji has been a member of the national leadership team of Pratham, one of India’s largest non-government organisations working in education. Educated in India and in England (Rhodes Scholar at Oxford University from 1981 to 1983), Rukmini Banerji completed her PhD at the University of Chicago in 1991. Her work with Pratham since 1996 has focused on designing and implementing large-scale programs for improving primary school student outcomes through collaborative partnerships with state governments and also working directly with village communities. Dr Banerji has also been the Director of ASER Centre, the autonomous research and assessment unit of Pratham. For 10 years, she has led the Annual Status of Education Report (ASER) effort, the largest annual study ever done by Indian citizens to monitor the status of schooling and learning in the country. ASER has been widely recognised for its innovative use of citizens’ participation in understanding and improving learning, assessment and the delivery of basic services. Dr Banerji is CEO of the Pratham Education Foundation.

Abstract

In countries such as India, impressive progress has been made in schooling. More than 95 per cent of children are now enrolled in school. But when we look at children’s learning, the situation is far from satisfactory. Available evidence suggests that in Grade 5, only about half of all enrolled children can read or do arithmetic expected at Grade 2 level. Faced with this crisis, how can assessment lead to effective instruction? ASER (Annual Status of Education Report) uses simple tools to assess the current level of children’s ability to read and to do arithmetic. Using this assessment, children are grouped for instruction by level rather than by grade. Appropriate methods and materials are used for each group to help children begin from where they are today and move to where they need to be. The ‘teaching-at-the-right-level’ approach has been found to be effective in many settings in India for building basic skills quickly. This ‘new thinking’ from India can provide large-scale solutions for the learning crisis faced in many parts of the developing world.
In India, impressive strides have been made over the last 25 years in providing schooling opportunities for all children. Even ten years ago, the progress towards universal enrolment was palpable. There were government primary schools in almost every habitation in the country and all available statistics for children in the elementary school age indicated that more than 90 per cent of children were already enrolled in school. For a country as thickly populated and diverse as India, this is no mean feat.

As the challenge of ‘schooling for all’ was being met, there was a sense from parents and teachers, planners, policymakers and practitioners, that what was happening in the schools was not satisfactory. A large fraction of India’s children who are in school today have parents who have either not had much schooling or are not literate. While such parents understand what it means to send a child to school, they often do not know what it means to support a child’s learning. Many assumptions underlie how our education system works. There has been a widespread belief that schooling will lead to learning and that more years of schooling is associated with students being able to do more. As far as education is concerned, parents, like governments, have been input focused and expenditure driven; both have believed that being able to spend more on education will solve most of the problems that children face. At least in developing countries, the fact that we may have to look more closely at the relationship between schooling and learning is only a very recent realisation.

Birth of a new approach

In 2005, in India, there was a new government in place at the federal level. A two per cent education levy had been collected from the general population to support universalisation of elementary education. In the public announcements of the new government, there seemed to be an interest in ‘translating outlays into outcomes’. Thanks to a combination of all these, 2005 seemed like a good time to take stock of how far we had come with schooling, and explore what needed to be done with children’s learning.

From 1996 onwards, Pratham had been working closely with children in low-income communities.1 On the ground too, we could see that almost all children were enrolled in school. And those who were not knew that they ought to be in school. However, most children from families that lived in slums or in villages seemed to need a lot of help coping with school work. Parents and teachers found it easier to support a child who makes good progress. For those who did not make the expected progress, whether they were left out (out of school) or left behind (in school), it was not clear what could be done. In Pratham we started with basic reading and arithmetic, and found that these two gave children, even as old as ten, a good foundation to build on and a confidence that allowed them to propel themselves forward in the education system (Banerji, Chavan & Rane, 2004).

If the goal is not only to have every child in school, but also every child learning well, what needs to be done? The first step should be a status report on where India was on schooling and learning. While there were plenty of data available on schooling in 2005, there was little information easily available for learning in primary grades. It was in this context that the idea for doing an Annual Status of Education Report (ASER) was born (Banerji, 2013).

What is ASER?

At its core, ASER is a simple exercise; a set of very basic reading tasks (recognise letters, common everyday words, a four-sentence simple paragraph of text at Grade 1 level, and an eight- to ten-line ‘story’ of text at Grade 2 level) and arithmetic tasks (recognise one- and two-digit numbers, a two-digit numerical subtraction problem with borrowing, a division task where three digits are to be divided by a one-digit number). These tasks are given to sampled children from age 5 to age 16. For every rural district in India, 30 villages are randomly picked from the census village list. In each village, 20 households are randomly sampled. All children in the age group 5 to 16 are assessed one-on-one on the ASER tasks described earlier. In each district, a local organisation or institution carries out this exercise. Each year the activities related to ASER start in August and the report is released in January of the following year. Every year on average ASER reaches close to 650 000 children in more than 16 000 villages across the country (Banerji, Bhattacharjea & Wadhwa, 2013).

The first ASER report was released in January 2006. It reported the status of schooling and learning for almost every rural district in India. The estimates for enrolment were very similar to the official figures. However, the estimates of children’s level of reading and arithmetic were a huge shock to us and to many more in the country. The report said that about half of all Indian children who have spent five years in school still could not read at a Grade 2 level. The arithmetic figures were similarly worrying. The reactions to these findings varied from the education establishment raising questions about the methodology to those who were convinced that it was time India moved from focusing on inputs to being much more outcome oriented (Banerji & Chavan, 2014).

---

1 Pratham is a non-government organisation working in children’s education and youth skillin in India. The mission is ‘every child in school and learning well’. Pratham has activities in 21 states in India. ASER Centre — the group that leads the Annual Status of Education Report, or ASER survey — is the autonomous research and evaluation arm of Pratham. See www.pratham.org and www.asercentre.org for more details.
राजू ताम का एक लड़का था। उसकी एक बड़ी बहन एवं एक छोटा भाई था। उसका भाई गाँव के पास के विद्यालय में पढ़ने जाता था। वह खुश रहता था। उसकी बहन बहुत अच्छी खिलाड़ी थी। उसे राती दोहरा लगाना अच्छा लगता था। वे तीनों रोज साम-साम मीठे-मस्ती करते थे।

गाना खुश मोसी पेपर झोला किला आण मोर

For Letters/Words: Ask the child to read any 5, out of which 4 must be correct.

For Numbers: Ask the child any 5 numbers, out of which 4 must be correct.

Ask the child to solve any 2 subtraction problems. Both must be correct.

Ask the child to solve any 1 division problem, which must be correct.
Every subsequent report reinforced these findings. Not only was the basic learning level of children in India low but it was also ‘stuck’ (Pritchett & Beatty, 2012) until 2010, after which there are signs of a downward decline. Further, the learning trajectories were flat suggesting that if children did not learn basic skills in the early grades, they were unlikely to gain them later.

Over the decade, other studies using different instruments and methods, including the government’s own periodic student achievement surveys, pointed to unsatisfactory levels of basic reading and arithmetic. Looking at trends over time using the cross-sectional data from every year, it is possible to follow the learning levels of cohorts as they move through the education system. The data suggest that the experience of each levels of cohorts as they move through the education system. The data suggest that the experience of each subsequent cohort is worse than that of the previous cohort — meaning that the reading levels in Grade 5 today are lower than they were in Grade 5 five or six years ago. India does not have any data other than ASER that look at basic learning levels on a nationwide scale, or annually or for a broad age range of children starting as early as age five.

The worrying results from ASER have led to a lot of discussion in India and abroad about children’s learning and how it can be measured. The ASER approach has also been scrutinised closely (ACER, 2014). Typically, large-scale assessments, especially the international measurements of student achievement, have originated in countries where education systems are “settled” — all children are in school and all schools are on official lists. Further, in many developed countries, the gaps are low between the curricular expectations, teachers’ ability to deliver what is expected, children’s performance and parents’ capacity to understand what their children should be gaining in primary grades. Thus, in such countries, pen-and-paper tests based on grade-level expectations make sense even in primary school.

Right from inception, the design and architecture of ASER has been very different from the usual large-scale assessments that are done in countries around the world. ASER takes into account specific characteristics of the Indian context. The foundations of ASER are built on these realities. Here are some features that makes ASER different:

- **Where:** In India, children go to many kinds of schools. There are government schools. There are also a wide variety of private schools, including low-cost schools, religious schools and non-formal schools. Not all of these are on official school lists. Further, attendance varies considerably across regions and types of families. In some states in India, daily attendance can be higher than 90 per cent, but there are also states where on average, only five or six out of every ten enrolled children are attending school on a given day. In this context, to get a representative sample of children for any assessment, there is no choice but to go to the household. Thus ASER goes to the child’s home and uses sampling at the household level to generate estimates of learning.
- **What:** Many children who are currently in school in India are far below their grade level. Even after several years of schooling, a large proportion of children may not have acquired foundational skills like reading, number knowledge or ability to undertake basic operations. Without acquiring these skills, children are unlikely to develop higher skills. Thus ASER decided to focus on a few basic skills for all children rather than on subject outcomes for each grade. The data from ASER indicate that even at higher grades, there are children who need help on basics.
- **How:** If reading is likely to be a problem in primary school, then reading skills need to be assessed. The easiest way to assess reading is to work with children one-on-one. Children who cannot read naturally cannot deal with written words and therefore cannot do pen-and-paper tests. If such children are made to do pen-and-paper tests, we cannot find out what to help them with.
- **Whom:** The majority of parents, especially mothers, of children who are currently in school, have themselves had little education or are not very literate. They understand the importance of ‘schooling’ but are often not confident about how to support their children’s ‘learning’. In such a situation, it is very important to de-mystify learning and to work towards taking parents along. While the assessment is going on in the community or in the household, it is very common that parents for the first time begin to understand what it is that their children ought to be picking up in school.

---

2 These are artificial cohorts based on repeated cross-sections.

3 Other than the points highlighted here, there are three other features of ASER that were designed keeping Indian conditions in mind. First, planning for elementary education in India is done at the district level. Hence data for learning need to be available at the district level. To be useful in the planning process, ASER estimates are generated at the district level (and then are aggregated to the state and national level). Second, India is a big country. To capture and sustain national attention and to represent all children, nation-wide coverage is needed. This is one of the reasons that ASER is done in every rural district across India. Third, to bring about a significant change in national priorities and mindsets, frequent and timely, current and reliable data are needed on children’s learning. ASER has been done annually for 10 years. Each ASER report is released like clockwork in mid-January and figures are available for the current year — that is, the year in which the data are collected.
Why: The simplicity of the ASER tool and the ASER assessment process is very useful in engaging a wide range of people in understanding where children are and in thinking about how to support them to make progress. At a micro level, parents’ involvement is essential for children’s progress in learning. At a macro level, widespread and large-scale participation and engagement by citizens is essential for changing policy and practice. In every district, a local organisation carried out the survey effort. The success of ‘schooling’ can be attributed to the fact parents and governments all understood the critical importance of children going to school. Each in their own way worked to make the goal of universal schooling happen. In India and in many other countries, we are at that point where the common understanding in society about ‘what learning looks like’ and ‘how to improve it’ needs to be built.

From assessment to action

One of the key features of the ASER tool is how easily it helps people see the problem and enables them to plan action.4 Here is how this happens. Imagine a village in India. Let us say we want to find out the status of schooling and learning of children in this village. Armed with the ASER tool, and helped by the villagers, we go to every house in the village and talk to the children and the families and request every child to spend a little bit of time with us doing the reading and the arithmetic tasks. (This can also be done in the school.) At the end of a few days, we put together the village report card. Let us say the exercise shows that there are 200 children of elementary school age in the village — all of whom are enrolled in the one government school in the village. So as far as schooling is concerned, everyone is in school. Now what about learning? The report card shows us the status for all children. For example, we find that 75 children are in Grades 3, 4 and 5 — 25 in each grade. The report of their reading results show a wide variation across grades and within grades.

Typically schools are organised by age and grade. (In this table you can see the grade-wise reading levels in each row). In India, the usual way to teach is for the grade-level teacher to use the prescribed grade-level textbook and teach from it. So if we look at Grade 3, we can see that only two children are actually reading at the Grade 2 level and a three children are at Grade 1 level. Out of 25 children, about 15 are still either only at ‘letter level’ or below. Teaching these children from the Grade 3 textbook is not effective or useful. They need to have activities and materials at their level to help them to grow. In fact, teaching the curriculum instead of teaching children usually means that many children get left behind even at a young age.

So what is to be done? Looking at the table again, we can think of another way to structure the teaching–learning activities for a few hours in these grades. Instead of teaching by grade and dealing with the wide range of levels in each grade, we can group children by levels (in the context of the table look at the columns). We now have five groups, with between 10 and 20 children in each group. If we had three teachers teaching one grade each in Grades 3, 4 and 5, now we can have them make three groups of children by level and teach them accordingly. (A possible grouping is suggested in the table.) There are some activities that all children do together such as listening to a story that is being read aloud and discussions around the story. Then within each group, using appropriate activities and materials, children work with their instructor. As a child makes progress, he or she can move into the next group. When sufficient children have moved, the groups and the instructors can be revised. Similar groupings can be made for arithmetic.

Table 1 Example report card: Children by grade and reading level

<table>
<thead>
<tr>
<th></th>
<th>Beginner</th>
<th>Can read letters but not words</th>
<th>Can read words but not sentences</th>
<th>Can read sentences at Grade 1 level but cannot as yet deal with a ‘story’</th>
<th>Can read a story at Grade 2 level</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 3</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>Grade 4</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>Grade 5</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>17</td>
<td>19</td>
<td>12</td>
<td>11</td>
<td>75</td>
</tr>
<tr>
<td>Possible groups</td>
<td>33</td>
<td>19</td>
<td></td>
<td>23</td>
<td></td>
<td>75</td>
</tr>
</tbody>
</table>

4 ASER tools are used in many circumstances. What is described here is one version of the action that is generated by the assessment.
Across many districts and states in India, schools and villages in India reorganise themselves for a few hours during the normal school day to carry out the teaching-at-the-right-level model. Two hours a day teaching in this way helps to accelerate children's basic reading and arithmetic skills and gets them ready quite soon to deal with the usual curriculum and textbooks for their grade.\(^5\)

The use of the simple ASER tool in this context helps not only to bring out and de-mystify the problem but also helps to design the pathway to a solution. Given the reality of children's learning levels in India, the huge backlog of basic skills in primary school and the way that teaching and learning is usually organised, moving in this way from assessment to action seems do-able by teachers and by community members. This ‘frugal innovation’ does not need many additional resources; it needs a reorganisation of time and existing resources. However what it does need is an understanding of the core problem and a strong commitment to seeking solutions.\(^6\)

Such work that starts with the ASER assessment provides an excellent practical illustration of what ‘new thinking’ about assessment can lead to. The ASER instruments are designed to establish clearly where individuals are in their reading. We then use this assessment information to target teaching and learning at an appropriate level (rather than on age or grade level) and continue to use different versions of the ASER tool to monitor the progress that individual children and groups of children make and to evaluate the effectiveness of this assessment to action model.

### Concluding thoughts

As interest in children’s learning and measuring outcomes increases locally and globally, more countries will have to undertake ‘new thinking’. They will have to make sure that the assessments that are coming into place have taken into account the needs of their children, and that the processes that will ensue are within the capabilities of local people to understand and to do. In countries where the culture of measurement — especially of measuring outcomes — is weak, designing, implementing and providing feedback may take time. Hence it is essential that each step is taken well and each step builds on the learnings of the previous step. Children's learning is an issue of national importance and therefore all assessment that is carried out must lead to concrete action.

---

\(^1\) This method, sometimes called ‘teaching at the right level’, is used wherever Pratham works directly with schools and communities. (In the 2013–15 period, Pratham worked in this way in 10,000 schools across India.) Such methods are also used when Pratham and state governments work together collaboratively in partnership programs. (In the 2014–15 school year, more than 5 million children were reached in such partnership programs.) See Pratham (2015). Pratham's direct work with communities and schools as well as partnership programs with governments have been evaluated using randomised control trials and found to be effective at improving basic reading and arithmetic. See Poverty Action Lab (2014).

\(^2\) For examples of specific cases where the district administration in a state in India adopted the ‘teaching-at-the-right-level’ model, see Banerji (2014) and Pratham Education Foundation (2013).

---

**References**

Assessment standards, ‘intentional alignment’, and dialogic inquiry

Claire Wyatt-Smith
Australian Catholic University

Claire Wyatt-Smith is Professor of Educational Assessment and Literacy and Institute Director of the Learning Sciences Institute Australia. Her research focuses on professional judgement, standards and moderation, with an aligned focus on curriculum and literacy education. Her publications address teachers’ assessment identities; large-scale standardised testing and its impact on learning; assessment adaptations for students with disabilities and assessment and new technologies. Claire’s research has attracted funding from the Australian Research Council and she has undertaken numerous government-funded large-scale longitudinal projects. Recent books include Assessment for Education: Standards, Judgement and Moderation (Sage, 2014) and Designing Assessment for Quality Learning (Springer, 2014), and she is the Series Editor for Springer’s assessment series, The Enabling Power of Assessment, through which she works with international networks of researchers focused on assessment and professional practice. Claire holds an advisory role in the National Council for Curriculum and Assessment, Ireland and a Professorial Teaching Fellowship, Hong Kong University, working in the Faculties of Education and Dentistry.

Abstract

Internationally, the policy move towards standards-aligned instruction is gaining momentum. In Australia, standards have assumed unprecedented prominence in education policy relating both to classroom practice and to teacher preparation and career progression. The move is also evident in the United States, where the lure of standards to inform improvement is clear: significant investment has been committed to longitudinal research to examine at state and district levels the desirable conditions for implementing standards, their impact on developing college- and career-ready teachers, and in turn, the impact on teacher instruction and student outcomes. Moves such as this are occurring in the absence of a general theoretical position that connects assessment and standards to meaning making. This paper argues for the pedagogical utility of standards understood as enabling critical inquiry into teaching and learning. The notion of ’intentional alignment’ of standards, curriculum and assessment is explored through two key questions: What do teachers bring to assessment? And: What is involved in a dialogic approach to assessment standards which values learners’ perspectives and their agency in improvement?
The call for assessment innovation and system reform has international reach (OECD, 2013). It is arguably more pressing today than in earlier periods for a range of reasons. Societal change, concerning levels of youth unemployment, and radical changes in workplaces are unmistakable, as is the increasing rapidity of change associated with new technologies. The calls in many countries for a flexible workforce are loud, with clear evidence that as technologies make an impact on the nature of work, they also make an impact on the capabilities, attributes and dispositions valued in workers. The continued rollout of new technologies and convergence possibilities mean that human communication is undergoing unprecedented change. What type of education is needed in these times, and in turn, what approaches to educational assessment are needed? Given that there is no prospect of futureproofing, as may have been an aspiration in former eras, and that the link between education and employment is now not as strong as it was in the 20th century for many, questions abound about the kinds of assessment that will benefit young people in preparing them for their futures.

Along with such changes are some troubling signs of youth disengagement from schooling, and the impacts on learning, wellbeing and longer-term employability that this can bring. This presentation seeks to take account of these developments. It presents the case for the role of assessment in learning to be understood as shared enterprise, with the learner and ‘quality’ at the centre. What becomes shared — modelled by the teacher and ‘tried on’ and developed over time by learners — is an assessment mindset.

The presentation starts with two questions that circle validity and that call for new thinking about and practices for assessment. More than two decades ago, Rowntree (1977) posed the question, How shall we know them? The emphasis in this era was on the teachers (as we) knowing students (as them). I want to start with the proposition that Rowntree’s question can be rephrased, as: How can students make themselves known? Accompanying this is a proposed move away from student voice to student agency in assessment, with a direct focus on broadening the students’ experience of assessment. Related to this, but often overlooked, is the need for students to learn how to recognise, critique and generate ‘good work’, developing and applying concepts of quality. Essentially, what is needed for this to occur is for the pedagogical utility of standards, together with judgment and quality, to come to centre stage. There needs to be a focus on the value of teachers’ and students’ engagement in dialogic inquiry into how learning occurs in classrooms (Nuttall, 2004).

The second question involves the notion of what is meant by ‘expectation’ as represented in standards, and further, how expectation is used to engage students in improvement efforts. Alignment of curriculum and assessment and clarity of expectations are identified as foundational in much assessment literature. However, the pedagogical use of standards connected to curriculum, teaching and learning has not been validated internationally by empirical research. While Assessment for Learning — with its core principles of student agency in their learning against clarity of expectations and appropriate feedback — has been taken up widely at national and regional levels in several countries including Australia, we have as yet relatively little evidence of the success of these policies (Baird, Hopfenbeck, Newton, Stobart, & Steen-Utheim, 2014; Black, 2015; Wyatt-Smith & Klenowski, 2014). Indeed, the notion of what is meant by ‘expectation’ as represented in standards lacks good empirical support. This paper calls for large-scale research to be undertaken and will introduce a study to this end.

Through the entry point of these questions, the presentation seeks to take discussion back to assessment foundations, and in particular, validity, and then to the present possibilities for action, for new thinking and professional practice in assessment. It gives an opportunity to connect assessment to large open questions about:

1. teacher assessment identity and the potential benefit of moving beyond the notion of assessment literacy
2. the role of teachers as assessment designers with a designer’s eye on and skill in developing students’ capabilities in goal-setting, their criterial knowledge1 and evaluative experience
3. the contribution of dialogic inquiry in the classroom as a means to support students’ meta-cognitive development including the assessment mindset discussed above
4. a move towards developing digital learning histories to build a richer picture of learning progression.

These four thematic lines lead to the ultimate question of why validity matters more than ever. A related intention is to reposition dialogue about ‘good teachers’ and ‘good teaching’ with implications for what it means to be ‘a good student’.

---

1 ‘Criterial knowledge’ refers to student knowledge of ‘criteria relevant to a fine performance on the task at hand’ and how to deploy this knowledge to inform on-task improvement strategies and self-monitoring (Wyatt-Smith, 2001, p. 118).
Assessment as professional capability

Assessment is now recognised as a key professional capability for teachers. Developing teachers’ assessment capability is recognised as a national priority for many countries including Canada, New Zealand, Australia, Ireland, Scotland and Japan. The Australian review of teacher education (TEMAG, 2015) recognises the need to lift teacher capability in assessment, in using standards and in using data, to improve student performance. Standard 5 of the Australian Professional Standards for Teachers addresses the collection, interpretation and use of assessment data to improve teaching practice (AITSL, 2014), including the provision of appropriate feedback to students. Section 4.2 of the Irish Professional Code of Conduct for Teachers focuses on the need for teachers to ‘maintain high standards of practice in relation to pupil/student learning, planning, monitoring, assessing, reporting and providing feedback’ (Teaching Council, 2012, p. 7).

Such emphases have been incorporated into education policy. The current Australian Curriculum was designed to meet the promise of the Melbourne Declaration on Educational Goals for Young Australians (MCEETYA, 2008, p. 5) linking schooling, equity and excellence with curriculum and expectations of ‘common high standards of achievement’, encouraging the Australian Curriculum, Assessment and Reporting Authority to work with state and territory systems to investigate ways to strengthen national consistency in application of standards (ACARA, 2012).

For illustrative purposes, and to broaden the focus beyond Australia, I draw on another national experience of curriculum and assessment reform in Ireland. The new Junior Cycle (NCCA, 2011) has been accompanied by a contentious shift in assessment policy whereby teachers assess and judge student work against stated standards and in using data, to improve student performance. Standard 5 of the Australian Professional Standards for Teachers addresses the collection, interpretation and use of assessment data to improve teaching practice (AITSL, 2014), including the provision of appropriate feedback to students. Section 4.2 of the Irish Professional Code of Conduct for Teachers focuses on the need for teachers to ‘maintain high standards of practice in relation to pupil/student learning, planning, monitoring, assessing, reporting and providing feedback’ (Teaching Council, 2012, p. 7).

Against this backdrop of changing curriculum and assessment contexts in Australia and elsewhere, the discussion commences with the issue of students’ intellectual engagement. This is taken as foundational in the new professional knowledge that locates assessment at the heart of pedagogy.

Connecting assessment, engagement and school-community partnerships

Dunleavy and Milton (2009) discussed the requirements for intellectual engagement. They identified the difficulties of isolating particular classroom practices that would be most effective in supporting it. While recognising these challenges, they proposed a set of common instructional ‘designs for learning that begin with the goal of intellectual engagement’ (p.13) that arguably have relevance to assessment that aims to trigger and sustain student engagement in learning. According to these writers, the designs:

- require high levels of student participation and provide time for in-depth work
- incorporate authentic assessment as a strategy that helps students set goals and assess their own learning
- use work that is relevant, interesting, and connects with students’ aspirations; is rigorous and allows students to think as ‘professionals’ and create professional-quality outcomes; is challenging and allows students to experience a sense of deep intellectual and emotional investment in learning; is built from diverse and improvable ideas; and is informed by the current state and growing knowledge bases of different subject disciplines
- promote students’ sense of ownership and responsibility for their own learning
- invite students to be co-designers of their learning in classrooms; support student voice and autonomy
- provide a high level of social support for learning and encourage students to take risks, ask questions, and make mistakes.
- foster collaboration and community building
- engage students in becoming literate with technologies as social networking-knowledge building tools
- connect students with opportunities to develop abilities in critical thinking, intellectual curiosity, reasoning, analysing, problem-solving and communicating
- bridge students’ experience of learning in and outside of school by exposing them to digital technologies in knowledge building environments (2009, pp. 13–14, emphasis added).

The above shows a general recognition that context matters, with support for authentic assessment and the role of students in setting goals and setting their own learning. They also highlight the relationship between school and community, and by extension, the world of work and community engagement. At the core of this relationship lies the traditional and powerful link between
assessment, on the one hand, and on the other, the control of curriculum; what students learn, and the tasks students are required to undertake both for learning and for assessment and grading. Bound up here are matters of teacher and school authority, and the potential for interdisciplinary individual or small-group student-initiated and led projects. From this perspective, human resources that include teachers, community members and industry could play a strengthened role in ensuring connections of in-school and outside-school learning and assessment, and in turn, post-school pathways.

It is useful here to distinguish between system and site validity (Freebody & Wyatt-Smith, 2004). Validity is taken to refer to what is assessed and how well this corresponds with the behaviour or construct that it is intended to assess (Harlen, 2004). In the case of ‘site validity’ it involves assessments that intend to assess the range of skills and knowledges that have been made available to learners in the classroom context or other sites. High ‘system validity’ involves assessments that intend to assess an often narrower range of skills and knowledges, regarded as essential by a government body or system.

Barriers to moving towards the strengthened focus on site validity come from current accountability requirements that rely heavily on large-scale standardised tests and thus work against the design-led assessment and instruction.

Australian research reports that teachers can experience the dual approach of assessment for learning purposes, and the prioritising of testing and test preparation, including for the National Assessment Program — Literacy and Numeracy, as presenting competing assessment demands. On the one hand, as McClay (2002) highlighted, there is increasing downward pressure to rehearse standardised testing conditions, to make students ‘test-savvy’, and to thereby demonstrate a type of quality assurance of learning and teaching. On the other hand, there are the imperatives to develop and implement assessments that have high ‘site validity’. Characteristic of such assessments, as noted elsewhere, are teachers’ efforts in connecting in-school and out-of-school knowledges. The aim routinely is for school activities to have touch points with contexts outside schooling (Cumming & Wyatt-Smith, 2001). The relationship between system validity and site validity has changed with the move in this country and others towards considerable investment in testing. As testing moves online, this relationship is likely to be impacted further.

Recent research (Cumming, Wyatt-Smith & Colbert, forthcoming; Ng, Wyatt-Smith & Bartlett, forthcoming) suggests that the potential benefits of standardised tests for improving learning are not being realised in classroom practice. It appears that this will continue to be the case until the links between testing and improvement efforts at system and school levels are more clearly articulated and better understood by teachers, students, parents and the wider community.

Related lines of inquiry

The connection between teachers’ assessment knowledge, curriculum standards, and teaching and learning is taken as being at the foundation of much work on assessment to improve learning but has not been validated in empirical practice. Further, a core tenet of assessment research on descriptive standards in standards-referenced systems is that the descriptions are guides as to what is required in students’ work to achieve a standard, and how it will be assessed. An additional tenet, which underpins much current writing, is that clarity of these assessment expectations is important for student learning, through goal setting and through feedback to students about the quality of the learning they have demonstrated and the gap that they may need to close to achieve a better learning outcome (ARG, 2002; Sadler, 1989). While such ‘assessment for learning’ has gained a hold in assessment policy worldwide (see, for example, MCEETYA, 2008), and is being widely implemented in different forms internationally, there is scant large-scale empirical evidence on how teachers and students work with stated assessment expectations and are able to use these to guide and improve both student learning and teaching practices (Black, 2015; Klenowski & Wyatt-Smith, 2014; Torrance, 2012; Wyatt-Smith, Klenowski, & Colbert, 2014), and further, how they build a shared assessment mindset.

The four thematic lines outlined earlier (identity, task design and standards, dialogic inquiry, and digital learning histories) are addressed in the presentation using data from a range of studies for illustrative purposes. While it is possible to treat each one separately, innovation lies in seeing them as a suite of connection points that inform teachers’ and students’ decisions about assessing what matters. At issue are both intellectual and relational synergies in developing the assessment culture of the classroom and the school more generally. They complement an approach to learning-centered task design and dialogic inquiry not only into what is learned, but also the cognitive and meta-cognitive processes underlying learning and performance.

The potential of rethinking assessment in these ways lies in reconsidering how a hallmark of ‘a good teacher’ could extend well beyond being recognised for the good grades that students achieve. Instead, the measure could be the success of teachers, leaders and school systems in developing students’ abilities to use existing knowledge, to generate new knowledge, and to think and deploy meta-cognitive knowledge. This is taken to include students’ insights into themselves as learners and how they learn, and moreover, how to apply knowledge and skills, and how to transfer and adapt them to be effective in new contexts, facing new problems and working in new collaborating teams. Students’ ability to meta-cognitively assess and adapt will promote their opportunities to contribute, to lead, and innovate in societies of the future.
References


Assessment in interactive learning environments

Michael Timms
Australian Council for Educational Research

Dr Michael Timms directs the Assessment and Psychometric Research Division at ACER, which develops high-quality assessments and conducts cutting-edge research in educational measurement. He is Chief Investigator in the Science of Learning Research Centre, in which ACER is a lead institution. He is a recognised leader in the development of innovative ways to assess students in electronic-learning environments. His research is widely published in peer-reviewed journals and he was awarded the 2013 Journal for Research on Science Teaching Award by the National Association for Research on Science Teaching.

Prior to joining ACER, Dr Timms was Associate Director of the Science, Technology, Engineering and Mathematics (STEM) Program at WestEd, a pre-eminent educational research and development organisation in the United States. He led large-scale research studies in STEM education, with special focus on computer-based assessment projects, especially through the SimScientists research program (www.simscientists.org). He has been involved in the development of two assessment frameworks for the US National Assessment of Educational Progress for which he was awarded the Paul Hood award for excellence in educational research at WestEd.

Dr Timms has experience in leading evaluation research projects for other educational research grant recipients, such as universities, and has managed large-scale item development projects across many content areas. He is knowledgeable about the education systems of Australia, the United States and the United Kingdom.

Dr Timms is an Associate Editor for the Australian Journal of Education.

Jason Lodge
The University of Melbourne

Dr Jason Lodge is a psychological scientist, research fellow in learning sciences and an experienced lecturer in psychology and higher education. Jason is part of the Melbourne Centre for the Study of Higher Education and is affiliated with the national Australian Research Council funded Science of Learning Research Centre at the University of Melbourne. He was also the recipient of the 2014 Australian Psychological Society Early Career Teaching Award. Jason’s areas of expertise are in cognitive science and educational psychology. His research focuses on the cognitive and emotional factors that influence student learning and the student experience in educational settings, particularly in digital learning environments.
Abstract

There is an increasing interest in using digital technologies to create interactive learning environments (ILEs) that both teach and assess student skills that are hard or impossible to assess using ‘static’ items such as traditional, multiple-choice questions. These interactive learning environments try to do two things simultaneously: firstly, to monitor the learning of the student in real time, providing feedback to help the student progress through the learning task; and secondly, to use the information gathered during the learning to make judgements about where the student is in learning of the topic. Essentially, ILEs draw upon the same source of data — the interactions of the student with the learning materials and embedded assessment tasks — to perform these measurements. To make these kinds of decisions, ILEs collect and analyse many variables; the complexity of these data demands the use of sophisticated assessment methods that differ from those used in traditional paper-and-pencil tests. The complexity of the ILEs also introduces challenges such as students becoming confused or failing to comprehend the feedback from the system.

Through reference to examples of ILEs, this session shows how assessment of learning takes place, how such assessment can provide valid and reliable measures, what we are learning about students’ use of the systems and how we are working to refine the systems of the future.
Much of the work in the design and implementation of interactive learning environments (ILEs) with embedded assessments has occurred in science education. The reason for this is that science education worldwide has increasingly focused on ensuring that students acquire not only the knowledge and conceptual understanding of the discipline, but the practices of science that follow the scientific method. Science practices typically include the application of such skills in the early grades as recognizing patterns and formulating answers to questions about the world. As they move on through the grades, students are expected to be able to gather, describe and use information about the natural world, and eventually to design experiments. This is being achieved through the use of digital materials that provide active and interactive learning scenarios in which students can apply what they have learned and engage in these science practices.

The United States has been particularly active in this area. The publication in 2012 of *A Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* and subsequent publication in 2013 of the *Next Generation Science Standards: For States, By States* called for a change in the way science is taught and assessed in the US. The framework advocated for a system of kindergarten to Year 12 science education that reflects the way that scientists work and think. It also called for research-based instruction that leads students to build conceptual understandings in science as they progress through their education. The framework emphasised an interweaving of the practices, crosscutting concepts and core ideas into the curriculum, instruction and assessment of the various disciplines of science. It used the term ‘three-dimensional science learning’ to refer to the integration of these dimensions. This three-dimensional science learning approach to science education also forms the basis of the *Next Generation Science Standards* (NGSS), which set out performance expectations that specify goals about what students should know and be able to do at each grade level.

To assist those who wish to design assessments of the NGSS, in 2014 the Committee on Developing Assessments of Science Proficiency in K–12 published *Developing Assessments for the Next Generation Science Standards* (National Research Council, 2014). The report refers to the need for classroom-based assessments that can form part of the overall assessment systems for science and this has led to many research projects in the US that have developed prototype systems.

In this paper, we show an example from the US of an ILE that both teaches and assesses simultaneously, and illustrate the kinds of measurement methods that are used to assess the learning that takes place. We also examine whether the assessment that takes place in this ILE can provide reliable measures. Finally, we discuss what has been learned to date about students’ use of such ILEs with embedded assessments and the implications for design of future systems.

### An example of an intelligent learning environment with embedded assessments

A genre of ILEs that has emerged is science learning modules based upon simulations of natural phenomena. Simulations have been chosen for science instruction because they offer some advantages. They can provide dynamic representations of spatial, temporal and causal phenomena in science systems. They can show things that are not directly observable, such as erosion over time, and they allow learners to explore and manipulate scenarios. Simulations also have the advantage of being able to present content in multiple representational forms, which has been shown in numerous studies to help students to build mental models of concepts and principles. In addition to having advantages for student learning, simulations offer advantages for assessment too. They offer the opportunity to design assessments of systems thinking, model-based reasoning and scientific inquiry which are seldom tapped in static, conventional tests. In other words, simulations offer opportunities to examine the learning process in addition to learning outcomes.

Another use of simulations in science is to provide virtual laboratory equipment that mimics what a student may find in a real science lab. The ChemVLab+ project ([www.chemvlab.org](http://www.chemvlab.org)), for example, provides chemistry activities that encourage students to solve authentic problems by designing experiments in a virtual chemistry lab (Davenport, Powers & Rafferty, 2014). Figure 1 shows two screenshots from an activity in the stoichiometry module. The top screenshot shows, on the right, the questions that students have to answer and, on the left, the virtual laboratory workbench in which they can select glassware, equipment and chemicals to conduct the procedures necessary to answer the questions. Rather than replacing classroom lab experiences, the ChemVLab+ activities are designed to replace lectures and traditional paper-and-pencil exercises. In the bottom screenshot of Figure 1, students are able to drag tiles that represent molecules to create a balanced chemical equation, a task that is not easy to do in paper-and-pencil tasks.

Each of the four activities in ChemVLab+ use a constraint-based modelling approach in which the errors that a student makes provide information about what the student knows and the kind of help the student needs. The data for these decisions are gathered from the student’s interactions with the activities and initiated
Figure 1 Screenshot of an activity from the stoichiometry module of the ChemVLab+ project. Top, students combine chemicals in the virtual lab to determine how the chemicals react. Bottom, students drag molecules to create a balanced chemical reaction. (http://chemvlab.org)
when the student clicks the ‘hint’ button or attempts to move on with incorrect responses. The learner receives tiered feedback in three levels. The student is first shown where errors have been made. Next, the student is told what scientific principles are relevant to the given problem. If the student continues to make errors, the hints provide the correct response with an explanation. Student proficiency is estimated using the number of errors they make on the concepts and skills that are the targets of instruction for the module. When a class has completed the activity, teachers can access reports that indicated areas of mastery and difficulty for students. See Figure 2 for an example of the summary report that teachers receive.

The question arises as to how reliable an assessment that is embedded in a complex learning environment can be. To test this, student response data from 1373 students from eleven US high schools that used the stoichiometry module has been modelled using item response modelling. The schools were a mix of urban, suburban and rural with a range of students from low to high socioeconomic status. Item response modelling is a method used to produce estimates of student ability in a wide range of assessments including large-scale assessments like Australia’s National Assessment Program — Literacy and Numeracy, for example. The data included dichotomous data points from across the four activities in the unit and scores from across the written responses in the four activities, which were scored by humans using rubrics. There were ten written response items: two items were scored 0, 1, 2 and eight items were scored 0, 1.

First, a unidimensional model that represented the whole of stoichiometry was applied to the dichotomous items and to the combined dichotomous and written response items. Two items (one dichotomous and one written response) that had psychometric characteristics outside the acceptable range were omitted from the analyses. The reliability (EAP) for the dichotomous items on their own was 0.93, and with the inclusion of the human-scored written items, the reliability increased to 0.95, a high level of reliability. A multidimensional analysis that produced student ability estimates for each of the seven content dimensions of stoichiometry was also conducted. The reliability estimates for each sub dimension are also good, demonstrating that the reports to teachers on what students know in these content dimensions are reliable to act upon. The reliability estimates are shown in Table 1.

Figure 2 Example of summary report for teachers (http://chemvlab.org)
Things to consider in designing embedded assessment systems in ILEs

To investigate the use of interactive assessments like those embedded in ILEs, De Boer et al. (2014) conducted a comparison of three modes of assessment for middle school students studying ecosystems. The study examined the comparative effectiveness of assessment tasks and test items presented in online modules that used either a static, active or interactive modality. A total of 1836 students used the assessments as part of normal classroom activities, taking assessments in the three different modalities on three consecutive days. The assessments tested key concepts about ecosystems and students’ ability to use inquiry skills in an ecosystems context. Figure 3 shows a comparison of the three types of static, active and interactive items and how they can be targeted to assess the same learning goals. The modalities varied in how much activity students saw on the screen and how much interaction and control students had in the testing environment. Also, the interactive modality allowed for some items in which the students were given the opportunity to apply their knowledge of the targeted learning goal by, for example, designing and running their own experiments. The equivalent item in the static and active modalities only asked students to evaluate and select correctly designed experiments.

De Boer et al. (2014) found that there were no significant differences in performance in two essentially identical items that appeared in all three modalities. However, in two different sets of items on which there were differences in the activity/interactivity of the items, students performed better on the static items than on the active and interactive modality items.

De Boer et al. suggest that there are two possible explanations: that the students had more difficulty with the content of the active and interactive items, or that they had difficulty with the technology. If content is the reason, then the interactive test may be tapping into more cognitively complex skills (for example, carrying out experiments compared to identifying a correct design). Alternatively, the active and interactive items may also require a higher degree of technical experience with interactive systems. In observations of some of the students using the interactive system, De Boer et al. noted that students did not always use the technology in the way it was intended that they should.

A number of students, for example, did not immediately understand how one feature that allowed them to inspect the graphs of results worked. Also, students did not go back to rerun simulations of the ecosystems but preferred to trust their memories of what they had just seen. This points to differences in the way that students may interact with the systems in which the assessments are embedded. This may be related to observations in other research on interactive learning environments.

By the time students are into their middle years of schooling, they have had much exposure to selected response assessment items, such as multiple-choice, in which they have to evaluate some choices and select the best answer. There is no level of confusion in such items, other than that caused by the content. This is not so as we move into complex interactions in ILEs where design decisions have been made about how a simulation may work within the limitations of the screen size and the interactions possible through a keyboard and mouse or a touch screen.

Table 1 Stoichiometry content dimensions

<table>
<thead>
<tr>
<th>Dimension</th>
<th># Items</th>
<th>EAP reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concentration</td>
<td>20</td>
<td>0.85</td>
</tr>
<tr>
<td>Unit conversion</td>
<td>34</td>
<td>0.92</td>
</tr>
<tr>
<td>Molar mass</td>
<td>22</td>
<td>0.84</td>
</tr>
<tr>
<td>Balanced reactions</td>
<td>22</td>
<td>0.87</td>
</tr>
<tr>
<td>Using stoichiometry</td>
<td>11</td>
<td>0.81</td>
</tr>
<tr>
<td>Significant figures</td>
<td>14</td>
<td>0.87</td>
</tr>
<tr>
<td>Experimentation</td>
<td>31</td>
<td>0.86</td>
</tr>
</tbody>
</table>
Figure 3 Comparison of an item set in three modalities. Items differ in activity and interactivity (De Boer et al., 2014).
One main advantage simulations in particular offer is insight into the way that students approach and work through different content. For example, Dalgarno, Kennedy and Bennett (2014) found that, when given a simulation on blood alcohol concentration, higher education students tended to either take a highly systematic approach or a haphazard and unsystematic approach to working through the simulation. The students who relied on a systematic approach to understand the material performed significantly better in post-tests than the unsystematic group. In this instance, there was a distinct advantage to taking a scientific and systematic approach to understanding the material that was reflected in the behaviour students demonstrated in the simulation. This behaviour was also evident in the data captured during their learning and could thus be assessed.

While a systematic approach is useful in understanding many scientific concepts, in other cases students need to develop insight about a concept that necessitates a different way of thinking about it. Counterintuitive concepts such as Newton’s second law provide one example of this issue. Students often need to go through some form of cognitive disequilibrium or confusion before they can reconcile the new, counterintuitive information and their intuitive experience of the world to achieve conceptual change. In a similar vein to differences in approach found by Dalgarno and colleagues, evidence that students are experiencing this confusion and achieving conceptual change can be collected and examined in ILEs (D’Mello et al., 2014). Therefore not only can the conceptual change process be monitored and assessed in ILEs, personalised feedback can be given to students at the exact point at which they need it.

In systems that use feedback we also see differences in how students use the available help and how they process it. For example, recognising the need for help is a metacognitive skill that requires students to monitor their own progress and understanding (Aleven & Koedinger, 2000). Student ability also is a factor that influences how students perform in ILEs. There is research to suggest that higher-ability learners do better within computer-mediated environments that allow for more learner control, compared to lower-ability students who do not (Recker & Pirolli, 1992). Also, those students with higher ability have been shown to be better at using help after errors, compared to their lower-ability peers (Wood & Wood, 1999). Mason and Bruning (2001) showed that students with low achievement levels perform better on both simple and complex tasks when feedback is immediate. However, students with high achievement levels perform better with delayed feedback, particularly on complex tasks. So, as we transition to interactive learning environments with embedded assessments that offer feedback, there are more design considerations to be made than in traditional assessments.

Conclusions

Interactive learning environments allow learners to engage in tasks that are able to simulate aspects of real-life scenarios and have consequently been used in a variety of science learning materials. They have been found to be useful in representing science phenomena that may be hard to observe in the classroom, such as an ecosystem, or to allow rapid and safe use of virtual laboratory equipment to conduct simulated experiments. Progress has been made in embedding assessment tasks into these learning environments which make use not only of students’ responses to traditional tasks such as selecting a correct response or typing in an answer, but also in monitoring their interaction with the components of the system. Embedded assessments that occur in real time can be evaluated immediately by the learning system and therefore can offer feedback to the learner, creating a strong formative assessment. They have also been used to provide summative feedback to the learner about their overall progress and to the teacher about the progress of the class as a whole or groups within the class. The assessments have also been shown to have acceptable psychometric qualities that confirm that they can produce reliable measures and that sound judgements can be made about learners using these methods. While progress has been made, it is still relatively early days for such interactive assessments and we are still learning that there are design choices in creating such assessments so that learners can derive learning benefits from them. Finally, we know that interactive assessments take a lot more time and effort to develop, and so we need to ensure that we use them for assessment of learning in areas that are hard or impossible to assess with active or static items.
References


Measuring what matters: Challenges and opportunities in assessing science proficiency

Abstract

A key challenge in shaping science learning for the future will be to develop new measures of learning that take into account what it means to be proficient in science (Pellegrino, 2013). The emergent view on proficiency, grounded in learning sciences research, emphasises using and applying knowledge in the context of disciplinary practice. Referred to as knowledge-in-use, this perspective on science proficiency is a centrepiece of the United States’ National Research Council’s (NRC) Framework for K–12 Science Education (NRC, 2012), embodied in the new US national standards (NGSS Lead States, 2013) and emphasised in the recently released NRC report on developing assessments to measure science proficiency (Pellegrino, Wilson, Koenig & Beatty, 2014). Central to this view is that disciplinary content — both disciplinary core ideas and crosscutting concepts — and practice should be integrated. This would mean that as students apply knowledge to make sense of phenomena and solve problems, they deepen their conceptual understanding of content as well as their understanding of how to do science. This paper provides a brief overview of a systematic and scalable approach for designing assessment items to measure student proficiency with new science learning goals that blend disciplinary core ideas and crosscutting concepts with practices. The assessment tasks are intended for formative use within classroom instruction. Drawing on prior research from assessment and curriculum design (for example, see DeBarger, Krajcik & Harris, 2014; DeBarger, Penuel & Harris, 2015), this paper presents such a design approach and considers implications of the overall work in this field.
Conceptual framework

The prior generation of US science standards (for example, NRC, 1996, 2000) treated content and inquiry as fairly separate strands of science learning, and assessments followed suit. In some respects, the form the standards took contributed to this separation: content standards stated what students should know, and inquiry standards stated what they should be able to do. Consequently, assessments separately measured the knowledge and practice components. The shift to integrating science practices with disciplinary core ideas and crosscutting concepts, as emphasised in new US standards, called the Next Generation Science Standards (NGSS; NGSS Lead States, 2013), is based upon studies of actual scientific practice and what we currently know about student learning (cf., recent synthesis reports such as Taking Science to School, NRC, 2007 and A Framework for K–12 Science Education, NRC, 2012). This research corpus points to the importance of integrating content (that is, disciplinary core ideas and crosscutting concepts) and practice by emphasising that rich science learning requires tight coupling of what students know and what they can do. This idea of science performance (NGSS Lead States, 2013) presents a different way of thinking about science proficiency in that disciplinary core ideas and crosscutting concepts serve as thinking tools that work together with scientific and engineering practices to enable learners to solve problems, reason with evidence, and make sense of phenomena (NRC, 2012). The idea of science performance also signifies that measuring proficiency solely as acquisition of core content knowledge is no longer sufficient.

Knowledge-in-use learning goals comprise the Next Generation Science Standards and are articulated as performance expectations. Each performance expectation combines a science or engineering practice, disciplinary core idea, and crosscutting concept into a single statement of what is to be assessed at the end of a grade level or grade band. It incorporates all three dimensions of knowledge in use by asking students to apply disciplinary knowledge and make connections to a crosscutting concept as they engage in a science or engineering practice. This integrated, knowledge-in-use perspective poses challenges for assessment design. At this time, there are very few examples of assessments that integrate science content and practices in a manner consistent with a knowledge-in-use perspective. There is tremendous need for this assessment design work, as assessment will play a central role in supporting implementation of the new directions in science education both in the US and internationally. Our approach to meeting this challenge uses principles of evidence-centred design (Almond, Steinberg & Mislevy, 2002). Evidence-centred design has been used in wide-ranging assessment design contexts, from the development of large-scale, high-stakes assessments to the design of classroom-based assessments and other proximal or close measurement instruments. Evidence-centred design emphasises the evidentiary base for specifying coherent, logical relationships among: (1) the learning goals that comprise the constructs to be measured (that is, the claims we want to make about what students know and can do); (2) the evidence in the form of observations, behaviours or performances that should reveal the target constructs; and (3) the features of tasks or situations that should elicit those behaviours or performances. The need for a principled approach to assessment design, such as evidence-centred design, was explicitly discussed in the United States’ National Research Council report on developing assessments aligned to the Next Generation Science Standards (Pellegrino et al., 2014).

Application of evidence-centred design to three-dimensional science assessment

Figure 1 provides an overview of our overall design process for constructing assessment tasks that align with the Next Generation Science Standards. Our process follows the logic of evidence-centred design and contains three distinct phases — unpacking (domain analysis), constructing an assessment argument (domain modelling), and task and rubric development. While the figure represents a linear process that begins with selecting performance expectations and unpacking the three dimensions, it is important to realise that the process is very iterative in nature. The step of specifying evidence statements, for example, has caused us to revisit and revise our learning performances and unpacking.

Domain analysis: Unpacking components of performance expectations

In evidence-centred design, domain analysis typically entails gathering substantive information about how knowledge is acquired and used in a domain such as physical or life science. A Framework for K–12 Science Education and the Next Generation Science Standards specify meaningful ways to integrate the content and practices to promote assessment of learning in each domain. The analyses of the domain inform the construction of learning performances that represent formative assessment opportunities to check in on student progress toward performance expectations.

Unpacking the disciplinary core ideas

In this phase of evidence-centred design, we first unpack core ideas associated with a cluster of Next Generation Science Standards performance expectations at a given grade level or grade band by elaborating the meaning
of key terms, defining expectations for understandings for the targeted student level, determining assessment boundaries for content knowledge; identifying background knowledge that is expected of students to develop a grade-level-appropriate understanding of a disciplinary core idea; and considering research-based problematic student ideas and misconceptions.

**Unpacking the science practices**

Our unpacking of the science practices involves consideration of the core components of the practice, intersections with other science practices and the evidence required to demonstrate the practice.

**Unpacking the crosscutting concepts**

Unpacking the crosscutting concepts involves identifying the important components and opportunities for intersections with the science practices and with the particular disciplinary core ideas that are the target of the assessment.

**Domain modelling: Specifying a knowledge-in-use assessment argument**

Leveraging the unpacking of science practices, crosscutting concepts and disciplinary core ideas, we then move toward specifying a knowledge-in-use assessment argument. In this step, we consider relationships among the claims we want to make about what students know and can do, evidence that would demonstrate competency with respect to these claims, and features of tasks to elicit the desired evidence (see Table 1). Our claims, evidence, and task features reflect a knowledge-in-use perspective in that we emphasise the application of core ideas and crosscutting concepts through engagement in a science practice. Each claim takes the form of what we refer to as a learning performance. Each learning performance clearly describes what we expect students to demonstrate to provide evidence that they have achieved an aspect of a performance expectation. To construct learning performances, we identify the key aspect(s) of a disciplinary core idea, practice and crosscutting concept from our unpacking work, to specify statements of what a student should be able to do. As such, learning performances integrate aspects of disciplinary core ideas, practices and crosscutting concepts, and are written to express knowledge in use. Learning performances, however, are of a smaller grain size than performance expectations. Together, a set of learning performances provides the detail needed to create a coherent and bundled set of assessment tasks that would provide evidence that students can use the knowledge aligned to a performance expectation or cluster of performance expectations. In this way, high-quality learning performances function in relation to other learning performances to identify ‘what it takes’ to make progress toward meeting a standard (for example, Next Generation Science Standards performance expectations). Learning performances are also helpful for teachers as they help to identify an important opportunity that teachers should attend to and assess before the end of an instructional unit.

Once a learning performance has been specified, we then express the evidence students need to demonstrate to show they have met the claim. This can be thought of as student behaviours or performances that provide evidence of attaining the learning performance. To complete our assessment argument and before we can write assessment tasks, we need to specify characteristic and variable task features. Characteristic task features describe the attributes that are common across all the tasks for a learning performance. For instance, one characteristic task feature is that all tasks need to provide a motivating context. Variable task features describe what features can vary across the tasks. For instance, the level of scaffolding is one example of a variable task feature. Table 1 presents a knowledge-in-use assessment argument for a claim integrating disciplinary content knowledge about structure and properties of matter and the crosscutting concept of patterns with scientific practice of constructing a scientific explanation.

**Developing tasks and rubrics**

The final phase of the design process involves using the information detailed in the assessment argument to develop actual assessment tasks that will be presented to students. The task design depends on the specification of the characteristics and variable task features and allows for assembly of multiple tasks within a ‘family’ where the variations among the tasks could readily reflect intended levels of challenge. The task design process also takes into account the forms of evidence needed to support the learning performance claim and the ways in which that evidence will be scored and evaluated for purposes of rubric development. Obviously, validation of our assumptions about the tasks depends on collecting various forms of empirical data from students under conditions where we have a reasonable set of assumptions of the prior opportunity to learn.

**Discussion and implications**

Our design approach provides a broadly accessible vision of how to design Next Generation Science Standards assessments and is a vehicle for documenting principled design decisions. The systematic process anchored in evidence-centred design allows us to create well-aligned tasks that are usable across varied classroom environments. Although we have focused our efforts to date on physical sciences disciplinary core ideas and only a subset of the scientific and engineering practices, our process should generalise to other core ideas, crosscutting concepts and practices.
Table 1 Knowledge-in-use assessment argument

<table>
<thead>
<tr>
<th>Learning performance (claim)</th>
<th>Additional knowledge, skills and abilities</th>
<th>Evidence required to demonstrate proficiency</th>
<th>Characteristic task features</th>
<th>Variable task features</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Students should be able to construct an explanation about how they determine substances are the same based upon characteristic properties</td>
<td>• Knowledge that some properties can be used to identify substances, and that these properties are called characteristic properties (e.g., density, melting point, boiling point)</td>
<td>• Written claim: statement that substances (e.g., Liquid A and B) are the same or are different</td>
<td>• Assessment is limited to analysis of the following characteristic properties: density, melting point, boiling point, solubility, flammability and odour</td>
<td>• Types of properties included as data/evidence</td>
</tr>
<tr>
<td></td>
<td>• Knowledge that temperature, volume, and mass cannot be used to identify substances and are not characteristic properties</td>
<td>• Stated evidence: identification of at least two characteristic properties to support claim</td>
<td>• The term ‘substance’ means a pure material (not a mixture of substances).</td>
<td>• State of matter of substances (i.e., solid, liquid, or gas state)</td>
</tr>
<tr>
<td></td>
<td>• Ability to identify patterns in data on physical properties of different substances</td>
<td>• Description of reasoning: statement that the same substance must have the same set of characteristic properties or that different substances have different characteristic properties</td>
<td>• Tasks provide data about characteristic properties of substances</td>
<td>• Inclusion of irrelevant data (e.g., non-characteristic properties)</td>
</tr>
<tr>
<td></td>
<td>• Ability to identify which data can be used as valid and appropriate evidence</td>
<td></td>
<td>• Tasks provide a motivating/authentic context</td>
<td>• Level of scaffolding to develop claim, evidence and reasoning</td>
</tr>
</tbody>
</table>
While our design approach has important advantages, challenges also exist. From a learning perspective, integrated assessment of key aspects of all three dimensions seems to be feasible and should provide insights into student achievement and its change over time with instruction. However, such an approach brings unique challenges from the perspective of measurement and interpretation of performance. A central question is whether rubrics should integrate the Next Generation Science Standards dimensions into a single score or separately evaluate aspects of performance for all the three dimensions. This involves issues related to ease of use and feasibility, including the extent to which each of the three performance components are separable and identifiable. Teachers will also need professional development on how to use these items in the classroom. Thus, creating models of how three-dimensional items can be used formatively in the classroom will be instrumental for effective classroom use.

We believe that our program of research and development will help to provide answers to critical questions related to the design and use of assessments of science knowledge in use. A critical need exists for research and development of high-quality assessments that align with the Next Generation Science Standards that express knowledge-in-use learning goals. More important, teachers need to be able to use these tasks in classrooms to provide themselves and students with information about progress towards meeting the performance expectations. Having exemplary formative assessments that integrate core disciplinary ideas, scientific and engineering practices and crosscutting concepts will be important to multiple stakeholders. Teachers, students, parents and school officials are interested in using high-quality assessments that provide information preparing students for university and career readiness in the fields of science, technology, engineering and maths. Assessment researchers need to better understand the design principles and psychometric properties of assessments that integrate core ideas, crosscutting concepts and science practices. Science education researchers want to use the assessments to better understand larger issues that widespread adoption of a three-dimensional learning perspective would entail, including developing and evaluating new science curricula. Science educators and policy-makers want assessments that help them to better understand students’ knowledge and abilities and also to inform changes in classroom instruction.

References


Acknowledgements

Project members contributing to this work include Chanyah Dahsah, Dan Damelin, Cynthia D’Angelo, Louis DiBello, Brian Gane, Christopher Harris, Joseph Krajcik, Jane Lee, and Kevin McShaneley. We also gratefully acknowledge Angela DeBarger from Lucas Education Research, Deborah Pellegrino, J.W., Wilson, M., Koenig, J., Beatty, A. (Eds.) (2014). Developing assessments for the Next Generation Science Standards. Washington, DC: National Academies Press.

Pellegrino, J.W., and Tiffany Leones and Reina Fuji from SRI International for their contributions to this research and development project. This paper is based upon work supported by the National Science Foundation under Grant Numbers DRL-1316903, 1316908, 1316874. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.
Conference papers
Tuesday 18 August
‘Hacking may be the catalyst for creative educational design and improvements to assessment.’
— Associate Professor Phillip Dawson
Stealth assessment in video games

Val Shute
Florida State University, USA

Val Shute is the Mack and Effie Campbell Tyner Endowed Professor in Education in the Department of Educational Psychology and Learning Systems at Florida State University. Before coming to FSU in 2007, she was a principal research scientist at Educational Testing Service, where she was involved with basic and applied research projects related to assessment, cognitive diagnosis, and learning from advanced instructional systems. Her general research interests hover around the design, development and evaluation of advanced systems to support learning, particularly related to 21st-century competencies. Her current research involves using games with stealth assessment to support learning of cognitive and non-cognitive knowledge, skills and dispositions. Her research has resulted in numerous grants, journal articles, books, chapters in edited books, a patent, and a couple of recent books, including Measuring and supporting learning in games: Stealth assessment (Shute & Ventura, The MIT Press, 2013) and Innovative assessment for the 21st century: Supporting educational needs (Shute & Becker, Springer-Verlag, 2010).

Abstract

Games can be powerful vehicles to support learning, but their success in education hinges on getting the assessment part right. In this presentation, I will explore how games can use stealth assessment to measure and support the learning of competencies critical for the future. I will discuss what stealth assessment is, why it is important, and how to develop and accomplish it. I will also provide examples within the context of a game called Physics Playground that I designed and developed with my team. I’ll share what has been learned by recent research on stealth assessments in games, including:
• Does stealth assessment provide valid and reliable estimates of students’ developing competencies, including qualitative understanding of physics, persistence, and creativity?
• Can students actually learn anything as a function of gameplay?
• Are games designed with stealth assessment capabilities still fun?
Preparing our kids to succeed in the future requires fresh thinking on how to design new kinds of assessments that overcome the limitations of traditional assessments, such as multiple-choice tests and self-report questionnaires, and also support learning. Traditional assessments are often too simplified, abstract, and decontextualised to suit current education needs. Alternatively, we can dynamically assess students in engaging, situated environments (like well-designed games) rather than having students fill in bubbles on a standardised test form. We can also provide immediate, ongoing feedback to support learning.

A century ago, traditional assessments were fine because a person who acquired basic reading, writing and maths skills was considered to be sufficiently literate. The goal was to prepare young people for production jobs, because 90 per cent of students were not expected to seek or hold professional careers. But when faced with highly technical and complex problems in today’s world, we need to re-examine the nature of educationally valuable skills. Except in rare cases, our current education system neither teaches nor assesses these new competencies, despite a growing body of research showing that skills and dispositions such as persistence, flexibility, creativity, self-efficacy, critical thinking, systems thinking, openness, problem-solving and teamwork (to name a few) can positively impact student academic achievement and other aspects of life.

Games, assessment and learning: A new approach

Increasingly, research shows that digital games can support learning. However, this is usually shown using pre-test–game–post-test designs, where the pre- and post-tests measure content knowledge. Such traditional assessments don’t capture and analyse the dynamic and complex performances that inform modern competencies. How can we both measure and enhance learning in real time? I believe that a performance-based approach to assessment is needed. The main assumptions underlying this new approach are that: (a) learning by doing (required in gameplay) improves learning processes and outcomes, (b) different types of learning and learner attributes may be verified and measured during gameplay, (c) strengths and weaknesses of the learner may be capitalised on and addressed, respectively, to improve learning, and (d) feedback can be used to further support student learning.

In a typical digital game, as players interact with the environment, the values of different game-specific variables change. For instance, getting injured in a battle reduces health, and finding treasure or other objects increases your inventory of goods. In addition, solving really hard problems in games permits players to gain rank or ‘level up’. One could say that these are all ‘assessments’ in games: of health, personal goods and rank. But now consider monitoring educationally relevant variables at different levels of granularity via games. In addition to checking health status, players could check their current levels of, for example, systems-thinking skill and teamwork, where each of these competencies is further broken down into constituent knowledge and skill elements (for example, teamwork may be broken down into cooperating, negotiating and influencing skills). If the values of those competencies got too low, the player would likely feel compelled to take action to boost them.

One main challenge for educators who want to employ or design games to assess and support learning is making valid inferences — about what the student knows, believes and can do — at any point in time, at various levels, and without disrupting the flow of the game. One way to increase the quality and utility of an assessment is to use evidence-centred design, which informs the design of valid assessments and yields real-time estimates of students’ competency levels across a range of knowledge and skills. Accurate information about the student can be used as the basis for delivering timely and targeted feedback. This information can also be used for presenting a new task or quest that is right at the cusp of the student’s skill level, in line with Csikszentmihalyi’s flow theory and Vygotsky’s zone of proximal development. Given the goal of using educational games to support learning, we need to ensure that the assessments are valid, reliable, and also pretty much invisible (to keep engagement intact). That’s where ‘stealth assessment’ comes in.

Overview of stealth assessment

Very simply, stealth assessment refers to evidence-based assessment that is woven directly and invisibly into the fabric of the learning or gaming environment. During gameplay, students naturally produce rich sequences of actions while performing complex tasks, drawing on the very skills or competencies that we want to assess. Evidence needed to assess the skills is thus provided by the players’ interactions with the game itself (that is, the processes of play). These can be contrasted with the product of an activity, which is the norm for assessment in educational environments.

By analysing a sequence of actions within a problem or quest (where each response or action provides incremental evidence about the current mastery of a specific fact, concept or skill), stealth assessments within game environments can infer what learners know and don’t know (or can and can’t do) at any point in time. Now, because we typically want to assess a whole cluster of skills and abilities from evidence coming from learners’ interactions within a game, methods for analysing the sequence of behaviours to infer these abilities are not as obvious. As suggested above, evidence-based stealth assessments can address these problems.
When assessment is seamlessly woven into the fabric of the learning or gaming environment so that it’s virtually invisible — blurring the distinction between learning and assessment — this is stealth assessment. It is intended to be invisible and ongoing, to support learning and to remove (or seriously reduce) test anxiety while not sacrificing validity and consistency. A good way to describe stealth assessment is with a metaphor. Consider the way that businesses were run before the onset of barcodes in the mid-1970s. Before barcodes, businesses had to close down once or twice a year to take inventory of their stock. But with the advent of automated checkout and barcodes for all items, businesses today have access to a continuous stream of information that can be used to monitor inventory and the flow of items. Not only can a business continue without interruption, but the information obtained is far richer than before, enabling stores to monitor trends and aggregate the data into various kinds of summaries, as well as to support real-time, just-in-time inventory management.

Now think about approaches to assessment in schools today. They are usually divorced from learning where the typical educational cycle is: Teach. Stop. Administer test. Repeat loop (with new content). But with stealth assessment, schools would no longer have to interrupt the normal instructional process at various times during the year to administer external tests to students. Instead, assessment would be continual and invisible to students, supporting real-time, just-in-time instruction. The remainder of this short paper will briefly describe evidence-centred design (which undergirds stealth assessment), and present a short example of a game that has three stealth assessments running within it.

Stealth assessment and evidence-centred design

Stealth assessment uses an assessment design framework referred to as ‘evidence-centred design’, formalised by Robert Mislevy, Linda Steinberg and Russell Almond in the late 1990s. In general, the primary purpose of any assessment is to collect information that will allow the assessor to make valid inferences about what people know, believe and can do, and to what degree (collectively referred to as ‘competencies’ in this paper). Accurate inferences of competency states support instructional decisions that can promote learning. Evidence-centred design defines a framework that consists of several conceptual and computational models that work in concert. The framework requires an assessor to: (a) define the claims to be made about learners’ competencies, (b) establish what constitutes valid evidence of the claim, and (c) determine the nature and form of tasks or situations that will elicit that evidence. Each of these models are now described.

Competency model. The first model in a good assessment addresses the question: What collection of knowledge, skills and other attributes should be assessed? Variables in the competency model describe the set of personal attributes on which inferences are based. The term student (or learner) model is used to mean an instantiated version of the competency model — like a profile or report card, only at a more refined grain size. Values in the learner model express the assessor’s current belief about the level on each variable within the learner’s competency model.

Evidence model. The second model is the evidence model which asks: What behaviours or performances should reveal those constructs identified and structured in the competency model? An evidence model expresses how the student’s interactions with and responses to a given problem constitute evidence about competency model variables. The evidence model attempts to answer two questions: (a) What behaviours or performances reveal targeted competencies; and (b) What’s the statistical connection between those behaviours and the competency model variable(s)? Basically, an evidence model lays out the argument about why and how observations in a given task situation (that is, student performance data) constitute evidence about competency model variables.

Task model. The third model addresses the kinds of tasks or situations that should be created to elicit those behaviours that comprise the evidence. A task model provides a framework for characterising and constructing situations with which a learner will interact to provide evidence about targeted aspects of knowledge or skill related to competencies.

As learners interact with tasks or problems during the solution process, they are providing a continuous stream of data that is analysed by the evidence model. The results of this analysis are data (such as scores) that are converted to probabilistic estimates of competency state, which are then passed on to the competency model which updates the claims about relevant competencies. In short, evidence-centred design provides a framework for developing assessment tasks that are explicitly linked to claims about personal competencies via an evidentiary chain (for example, valid arguments that serve to connect task performance to competency estimates), and are thus valid for their intended purposes.

Brief example of stealth assessment

Physics Playground is the name of a computer-based game with two-dimensional physics simulations for gravity, mass, potential and kinetic energy, transfer of momentum, and so on. The goal of all 75 levels in the game is to guide a green ball over to hit a red balloon.
Everything in the game obeys the basic rules of physics. Using the mouse, players draw coloured objects on the screen, which ‘come to life’ when drawn. These objects apply Newtonian mechanics to get the ball to balloon and they include simple machines such as levers, ramps, pendulums and springboards.

Three stealth assessments are coded deeply into the game: measuring creativity, conscientiousness, and qualitative physics understanding. Competency and evidence models were created for each of the constructs. This entailed, per construct, about a 10- to 12-month literature review, then structuring the main competency variables into a model. Evidence was defined as the things a person did in the game that would provide information about particular competency variables. Task models provided a blueprint for creating all of the levels in the game. Levels increased in difficulty across the seven different playgrounds, and each level focused on eliciting evidence related to particular aspects of Newton’s laws of motion.

For instance, conscientiousness was modelled with four main facets: persistence, perfectionism, organisation, and carefulness. For the persistence facet, we defined a set of observables (behaviours in the game providing relevant evidence) that included the following: time spent on unsolved levels, number of restarts of a level, and number of revisits to unsolved levels. The game automatically tallies this information in log files that are then analysed by the stealth assessment machinery. The difference between answering self-report questions about persistence (for example, ‘I always try my hardest’) and actually exerting substantial effort when trying to solve a hard problem in the game is a clear example of the expression: Actions speak louder than words. And they do.

**Conclusion**

Our current capacity to assess students is often limited as it is based on a relatively small number of test items. As we move to a seamless assessment model, we will be able to more accurately assess students since we will have access to a much broader collection of students’ learning data. More accurate assessments enable us to better support student learning across a range of important educational areas.

**References**


Abstract

Hackers exploit weaknesses in a system to achieve their own goals. In this paper I argue that hacking presents a significant threat to the growing world of online assessment. This threat needs to be addressed through a variety of means; technological anti-hacking approaches will not be sufficient. The most effective ways to prevent hacking may be changes to the assessment tasks themselves to make hacking less tempting; these approaches also have a range of positive side effects in terms of authenticity, transparency of criteria, and ensuring tasks involve work beyond the exam. I conclude with a brief exploration of the ways that teachers may also hack assessment systems.
The promise of online assessment

Vast bodies of research indicate that when used appropriately, educational technology can improve learning outcomes for students (Means, Toyama, Murphy, Bakia & Jones, 2010; Tamim, Bernard, Borokhovski, Abrami & Schmid, 2011). Benefits from educational technology are greatest when we adapt curriculum, instruction and assessment to take advantage of the affordances of technology.

Assessment can be adapted to use technology in a variety of ways. Student learning and performance can be improved through automatic feedback on an exam, or allowing typing instead of writing (Butler & Roediger, 2008; Charman, 2014; Mogey, Cowan, Paterson & Purcell, 2012; Mogey & Hartley, 2013). Student judgement can be improved through formative self- or peer-assessment procedures, which are made more efficient thorough online systems (Li et al., 2015). Examinations can be made more authentic by incorporating rich computer-based tasks (Hillier & Fluck, 2013). Technology even enables a vast array of new assessment types, ranging from social media tasks to high-fidelity simulations.

Threats to online assessment

In addition to providing additional affordances for learning, technology-supported assessments also provide potential affordances for cheating. Existing research suggests that an unsettlingly high proportion of students have engaged in copy-paste plagiarism, with one 2008 study finding almost three in five students copy-pasting without citing (Selwyn, 2008). In response an arms race has developed around anti-plagiarism ‘text matching’ software such as Turnitin, which compares student work against a database of sources. Cheating students have adapted their practices, and now employ a range of clever strategies like running their copy-pasted sections through translation engines like Google Translate or Babelfish (Jones & Sheridan, 2014). In addition to assisting do-it-yourself plagiarists, educational technology has also supported the logistics of pay-for plagiarism, with essays available made to order.

Although online plagiarism has received substantial attention, the online underbelly of assessment hacking has received little mainstream scrutiny. Unfortunately this lack of awareness hides real threats to assessment integrity. In another paper (Dawson, 2015) I document several ‘proof of concept’ hacks on a particular type of electronic assessment system:

Bring-your-own-device electronic examinations (BYOD e-exams) are a relatively new type of assessment where students sit an in-person exam under invigilated conditions with their own laptop. Special software restricts student access to prohibited computer functions and files, and provides access to any resources or software the examiner approves. In this study, the decades-old computer security principle that ‘software security depends on hardware security’ is applied to a range of BYOD e-exam tools. Five potential hacks are examined, four of which are confirmed to work against at least one BYOD e-exam tool. The consequences of these hacks are significant, ranging from removal of the exam paper from the venue through to receiving live assistance from an outside expert.

Potential mitigation strategies are proposed; however, these are unlikely to completely protect the integrity of BYOD e-exams. Educational institutions are urged to balance the additional affordances of BYOD e-exams for examiners against the potential affordances for cheaters.

That paper has a troubling finding: even with in-person invigilation it is possible to circumvent all of the security features of some assessment software. Any assessment conducted on student-owned hardware is in theory vulnerable to similar sorts of hacks.

How can we deal with assessment hacking?

One possible approach to this problem is to do nothing, in the hopes that hacking remains a niche or hidden issue. However several of the attacks I present in that paper could be easily packaged up by one crafty student and shared or sold to others. In the parallel world of computer game hacking, this is the approach taken by gamers who want an unfair advantage.

Another approach to dealing with hacking is to invest heavily in clever security measures to counter the threat posed by hackers. This is the approach taken in the computer game hacking world, where intrusive software is installed alongside games to monitor for cheating and instantly ban offenders. Despite ever-increasing anti-cheating measures, hackers still identify new exploits on a regular basis, which sell for substantial sums online. In the online gaming world it appears that fighting hackers through technical means is still only partially successful.

An alternative solution to this problem may lie in educational rather than technological changes. If we start from the position that all of our assessment is vulnerable to hacking, what can we do to design tasks that still mostly achieve their purposes — even when hacked?

One of the threats posed by assessment hacking is that it may transform an examination from ‘closed-book’ to ‘open-book’, or even ‘open-book, open-web’ (Williams...
Hacking also threatens to reveal the marking logic that sits behind electronic assessment, which ranges from answers to multiple-choice questions, to intelligent scoring of written responses. Educational workarounds to this sort of threat may require us to move away from some task types entirely. They may also force us to make our marking criteria more transparent for automatically marked tasks.

Hackers can also make identities of those involved in assessment more difficult to verify, through impersonation or unauthorised collusion. In my own work I have been able to hack around secure systems and allow a Skype call or instant messaging chat to run in the background. These hacks challenge assessment designers to consider what they can ask of students that is uniquely theirs. So the threat of hacking may lead to more tasks that incorporate evidence of students’ work across a variety of verifiable situations over time.

Can hacking improve assessment?

Some of the adaptations required to combat hacking may result in assessment that is more authentic, transparent and sustained. But beyond changes to combat hacking, we can also think of hacking as a metaphor that can be applied to the process of assessment improvement.

In a recently completed Office for Learning and Teaching project (Dawson et al., 2014) we interviewed 33 university teachers about how they make changes to their assessment tasks. Several spoke about creatively interpreting the rules that surround assessment processes. Taking hacking as a metaphor, there is tentative evidence in our data that these teachers “hacked” around bureaucracy and complexity, in order to implement changes to their assessment.

Hacking is thus a powerful force in assessment, and one that will be very difficult to eliminate. However through creative educational design, hacking may be the catalyst for improvements to assessment.

References


The move to NAPLAN online: The advantages and the road ahead

Dr Rabinowitz has joined ACARA after serving as Senior Program Director of WestEd’s Assessment & Standards Development Services, a program of national scope with more than 100 staff throughout the USA. In that role, he served as Director of the national Center on Standards and Assessments Implementation and the Smarter Balanced Assessment Consortium Project Management Partner.

Dr Rabinowitz has consulted extensively on standards, assessment and school/educator accountability issues with researchers, policymakers and assessment staff at national, state and district levels in the USA. Through his involvement in more than a dozen state and national technical advisory committees and the role he served in the Center on Standards and Assessments Implementation, Dr Rabinowitz is very familiar with the standards and assessment system challenges and constraints faced by states and districts. He also has experience in supporting states as they design and implement new standards, assessment and accountability systems.

Dr Rabinowitz was also a member of the Common Core State Standards national validation committee. Prior to joining WestEd, Dr Rabinowitz served as State Assessment Director for the New Jersey Department of Education.

Dr Rabinowitz received his PhD in Educational Psychology and Statistics from the State University of New York at Albany, New York, USA.

Abstract

The task of developing and delivering the National Assessment Program — Literacy and Numeracy (NAPLAN) online presents enormous challenges. Nonetheless, the benefits of taking this on are well worth the efforts. With a tailored test design, NAPLAN online will provide a better measure of the Australian Curriculum, more precise results and a significantly faster turnaround of those results. NAPLAN will begin its delivery online in 2017, with an opt-in period until 2019. To ensure readiness, a large amount of research is being conducted through the Australian Curriculum, Assessment and Reporting Authority to inform the development process. This presentation will address the advantages of moving online, as well as outline several key research studies to be conducted prior to the test’s implementation, including:

- a trialling study to ensure all item types (traditional and innovative) work in the online environment
- a device-effect study to ensure no disadvantage to students regardless of whether they use laptops or tablets
- a fonts and readability study, to ensure the best layout for reading test (passages and items)
- an accessibility study to ensure the test is fair for all students, including students with disabilities, students in remote areas, and so on
- an autoscoring study, to demonstrate that computers can score NAPLAN writing as well as teachers.
With discussant Barry McGaw
McGaw Group

Professor Barry McGaw is Managing Director of McGaw Group and Honorary Professorial Fellow at the University of Melbourne. Professor McGaw has previously held roles as the Chair of the Board of the Australian Curriculum, Assessment and Reporting Authority, the Director for Education for the Organisation for Economic Cooperation and Development, and Executive Director of the Australian Council for Educational Research.
Translating rich learning assessments into certified results and university selection devices

Dr Matters recently reviewed senior assessment and tertiary entrance processes for the Queensland Government with co-investigator Professor Geoff Masters, culminating in the report, *Redesigning the secondary–tertiary interface*.

Her many other published journal articles, conference papers and books relate to statistical and social moderation, validity and reliability issues associated with test format, test-taking behaviour, criteria- and standards-based assessment, identifying and testing cross-curriculum skills, the alignment of curriculum, pedagogy and assessment, and the management of innovation.

Abstract

There are challenges in designing a set of high-quality processes in senior assessment and tertiary entrance that meet the needs of future senior secondary school students and future users of the certified results of learning assessments. Assessment and selection arrangements should look to the future rather than backwards to arrangements that might have existed in the past or that presently operate, unexamined, in other places. Teachers need to be convinced that the richness of students’ learning assessments will not be lost or transmogrified in any new processes for grading or ranking. A set of principles should guide the design of a new system — a set that gives pre-eminence to, but goes beyond, validity and reliability. This paper introduces the principles that guided deliberations in the recent review of senior assessment and tertiary entrance in Queensland, and describes, in simple terms, the design features of a new system based on the review’s recommendations.
Notes to the reader

This short paper incorporates but a small part of the 265-page review report, Redesigning the secondary–tertiary interface (Matters & Masters, 2014), available at http://www.acer.edu.au/queenslandreview

The terms of reference of the review can be found at www.acer.edu.au/queenslandreview/Review of Senior Assessment and Reporting and Tertiary Entrance Terms of Reference

The first person ‘we’ in this paper refers to Geoff Masters and Gabrielle Matters, the reviewers.

Until ACER’s 2014 report, the most recent review of tertiary entrance in Queensland had been in 1990.

For those unfamiliar with the Queensland system as it currently operates, a simple description can be found in Paper 1 in Volume 2 of the review report.

Queensland’s system of senior assessment and tertiary entrance, commonly referred to as the ‘OP system’, was established in 1992. The Overall Position (OP) is the primary selection device for Year 12 completers seeking entry to universities in Queensland. It is a rank order from 1 (highest) to 25 based on students’ overall academic achievement as measured by a combination of results across a student’s different subjects.

The ‘OP system’ covers more than tertiary entrance and the OP. The OP exists in the zone between school and university in which selection decisions are made; the OP system covers senior assessment as well as tertiary entrance.

Senior assessment in Queensland is school-based and externally moderated. There are no external examinations. Senior subject results are based exclusively on assessments (typically four to six) devised and marked by teachers in schools with reference to standards set down in subject syllabuses. The moderation model, designed to achieve comparability of standards, is consensus moderation, a form of social moderation that uses expert review panels at district and state levels. Senior subject results are certified as one of five levels of achievement (from Very High Achievement to Very Limited Achievement).

The Queensland Core Skills Test, a cross-curriculum test, is used to enable scaling of subject results in the calculation of OPs. Scaling is necessary before results in different subjects are aggregated because levels of achievement are not comparable across subjects. The Queensland Curriculum and Assessment Authority generates OPs and provides them to the Queensland Tertiary Admissions Centre, a company formed by the universities.

Review processes

Key aspects of senior assessment that the ACER reviewers (‘we’) examined were: Comparability, Moderation, Assessment instruments, and exit Levels of Achievement. Key aspects of tertiary entrance that we examined were: Overall Position (OP) and Field Position (FP), the Queensland Tertiary Admissions Centre (QTAC) Selection Rank, the Queensland Core Skills (QCS) Test, and the Australian Tertiary Admission Rank (ATAR).

Thousands of stakeholders and interested parties were involved in the review process: more than 2200 responses to a survey, nearly 100 formal submissions, four significant forums involving almost 300 key stakeholders and interested parties, and approximately 50 meetings of key stakeholders and their constituents with the reviewers. Thus we had many opportunities to gain insights into the way people were thinking about the OP system and to share our deliberations with them.

We made our own observations and undertook our own research, drew on our own knowledge and experience, built theories and tested out our findings with key stakeholder organisations, interested parties, technical experts, and colleagues in Australia and overseas who are influential in the fields of educational assessment, principles and practice and tertiary selection. We also sought counsel from our international consultant, Dr Peter Hill, who is renowned in education circles. We paid particular attention to two pieces of work commissioned for the review: Professor Claire Wyatt-Smith’s research into standards, teacher judgement and the operation of review panels as part of moderation; and Dr Reg Allen’s analysis of the strengths and weaknesses of the OP system today.

List of acronyms

ACER Australian Council for Educational Research
ATAR Australian Tertiary Admission Rank
FP Field Position
OP Overall Position
QCAA Queensland Curriculum and Assessment Authority
QCS Queensland Core Skills
QTAC Queensland Tertiary Admissions Centre
Outcomes of the review process

We identified three general areas in which we believe change is required in senior assessment processes for subjects approved by the Queensland Curriculum and Assessment Authority (QCAA) that can count towards tertiary selection indices:

- assessment activities
- assessing student performance
- moderation.

We identified three general areas in which we believe change is required in tertiary selection processes for Year 12 completers:

- use of a single rank order (OP, ATAR)
- separation of responsibilities for certification and selection
- transparency of procedures to those most affected by them.

We concluded that the current OP system, which has served Queensland well for more than 20 years, no longer functions as originally intended and is reaching the end of its usefulness. We recommended that it be retired and the secondary–tertiary interface redesigned.

The centrepiece of a redesigned system is a new Subject Result. We envisage that schools, through the QCAA, would produce valid, reliable, credible, stand-alone Subject Results for certification purposes on a fine scale (possibly 60 points) in place of five levels of achievement (Very High Achievement to Very Limited Achievement), and universities would use those results in fair, transparent and efficient ways as the basis for selecting students into their courses — most likely in the form of an ATAR.

As well as continuing to manage the processes for receiving and processing applications for the majority of undergraduate courses at Queensland universities, the QTAC would also devise any indicators required by the universities (such as an ATAR). In other words, the universities would no longer expect the school sector to rank their applicants for them.

Although we documented the weaknesses in an ATAR we understand why universities are committed to it — a national scale and an administratively simple selection procedure. Nevertheless, a 2000-point scale is untenable as that level of precision is not supported by the nature of the input data.

Furthermore, any rank ordering of students (OP, ATAR) is going to progressively break down over the next 15 years or so, as the basis on which rank ordering is built breaks down (for example, single cohorts of students all finishing their studies together as a group) and as the curiously Australian practice of aggregating scaled subject results is challenged. It is at the national level that the related discussion should occur.

For now, we recommended the introduction of prerequisites for high-demand courses, a reduction in the number of subjects that count towards a rank order, and a method for incorporating vocational education and training and other learnings into the calculation of rank orders.

Much of the rich information about student learning that is presently captured in school assessments is lost because of the coarseness of the reporting scale. There is a price to be paid, however, for a finer scale (say 60 points) — validity and reliability have to be enhanced. And so the proposed new design includes a prescribed assessment package, a simpler mechanism for marking student work, a revamped moderation model, and the addition of an external assessment (up to 50 per cent of the Subject Result).

A later section in this paper, ‘Underpinning principles’, relates our deliberations to recommendations.

Report to government

An interim report was provided to the state Minister for Education, Training and Employment in May 2014, a draft of the report uploaded to the ACER website in September 2014, and the final report submitted to the Minister in October 2014. There are 23 recommendations: seven on tertiary entrance, eleven on senior assessment and reporting, and five on implementation.

Government response

In January 2015, the Queensland Government released its draft response to the review report. The draft response provided in-principle support for a number of key proposals outlined in the review including the retirement of the current OP system. The draft government response was then subjected to further consultation with key education stakeholders and the broader Queensland community. Consultation continued until the end of March 2015 via an online survey and written submissions. The results of consultation will inform the development of a final Queensland Government response for release in mid-2015. There was a state election at the end of January 2015, which resulted in a change of government.

That the system’s weaknesses have been identified means that change is important and necessary. The nature of the weaknesses and the fact that those weaknesses are now in the public domain means that change cannot be delayed.

Observations and comments

In this short paper it is not possible to present findings formally. A collection of comments and observations is presented in Table 1. Even though the issues are interrelated, the comments and observations are presented in clusters. By the very nature of a review, critical comments prevail rather than comments about the strengths of the present system.

1 There are many other curriculum offerings but it is Authority subjects that count in the calculation of OPs. Elsewhere they would be recognised as tertiary entrance subjects.
### Table 1 Observations of existing Queensland tertiary entrance system

#### About the OP system
- Little or no support either among schools or universities for the OP system
- Different expressions of concern between and within the secondary and tertiary systems
- OP not aligned with the more diverse ways of completing senior secondary studies
- Almost half of Year 12 applicants are judged on criteria other than the OP; most notably a ranking of OP-ineligible students that does not take account of differences in subject difficulty or subject-group enrolments, thus creating an unfair binary system
- Lack of understanding of all aspects of the system by people at all levels of the secondary and tertiary sectors and the wider education community
- Anecdotes and perceptions abound
- Gaming by schools – for example, encouraging students to become ineligible for an OP, manipulating Subject Achievement Indicators, in the mistaken belief that gaming works in the ways intended
- Schools use OP results for marketing purposes

#### Discrimination
- OPs in 25 bands do not differentiate among applicants to high-demand courses so universities seek a finer scale
- Field Positions (up to five) are no longer useful in discriminating between students with the same OP
- There are only five grades for individual QCS results; not useful in discriminating between students with the same OP, especially at the top where it is needed
- The finer scale that the majority of universities desire is the national scale, an ATAR (30.00 to 99.95 in intervals of .05)
- The level of precision in an ATAR is not supported by the nature of the input data
- Rankings (such as OP and ATAR) are administratively simple for universities and QTAC

#### Lack of transparency in selection
- The original model of step-wise decision-making for selecting students (Overall Position, Filed Position, QCS grade, Level of Achievement, other admissible information) has changed on an ad hoc basis to include QCS percentiles and an ATAR (schools and students generally do not know this)
- The Queensland Curriculum and Assessment Authority’s calculation of an ATAR is based on data not generated for that purpose

#### Universities
- Demand-driven system means that hardly anybody misses out (less than 1.6 per cent of Year 12 applicants — a few hundred out of tens of thousands)
- Same ranking for students for all courses — but ranking is really only needed for high-demand courses to break ties between applicants
- Prerequisites are worth considering for high-demand courses (for example, taking highest-level mathematics subjects in secondary school to apply for engineering degrees at universities)
- OP is based on any combination of five Authority subjects out of approximately 50 subjects on offer — combinations can affect the rank order
- ATAR was an unknown species to the school sector before the review
Underpinning principles

Before we started our investigation we established the principles that would underpin our deliberations. At each stage of our thinking we reconciled our proposals against those principles.

**Principle 1 - Validity, reliability, utility**

Assessments of student attainment must provide valid, reliable and meaningful information about what individuals know, understand and can do, and how well, upon completion of Year 12.

**Implications of Principle 1**

- The purpose of certification is to confirm publicly students’ attainment levels upon completion of Year 12.
- Assessments of student attainment should be recorded on certificates in a form that is meaningful to students, their parents and schools, and useful to universities, employers and other users.
- Indicators of student attainment must be appropriate to a range of curriculum intentions, accurate and comparable across schools.
- Assessments of student attainment should stand alone and be independent of how they might subsequently be used.

2 Assessment processes are valid to the extent they provide information about the range of knowledge, skills and attributes identified in the senior curriculum. Assessment processes are reliable to the extent they provide accurate information about students’ levels of achievement comparable across students and schools.
It is also desirable that senior secondary assessment processes:

- promote high-quality teaching and learning in the senior secondary school, recognise the centrality of learning and reject anything that detracts from student learning
- have a futures orientation — assessment systems with a futures orientation are appropriate to the 21st century; recognise that curriculum priorities are changing; recognise that ways of assessment and learning are changing (responding to the role of technologies in teaching and learning); look to the future not the past; and are able to adapt speedily to changing circumstances
- are fair — that is, objective in the sense of not depending on who is doing the assessing.

Recommendations to enhance validity, reliability
1. Maintain and revitalise school-based assessment.
2. Add an external assessment (at least in some subjects).
3. Prescribe types of assessments to be undertaken and the conditions under which these assessments will occur.
4. Add results of school assessments and an external assessment to give an overall result for certification. However, the school assessment would not be statistically moderated against the external assessment.
5. Devise a new moderation model that involves endorsement of assessments before they are undertaken and confirms the attainment levels (marks) of students on those assessments, one at a time, over the course of study.

Recommendations to enhance usability
6. Devise a new way of describing performance against criteria, which is useful for arriving at Subject Results and for communicating those results to users.
7. Certify Subject Results on a finer scale (than at present) — say 60 points.

Principle 2 — Separation of responsibilities for senior certification and tertiary selection

Universities should take complete responsibility not only for deciding how their future students are to be selected (from the pool of Year 12 completers who seek admission) but also for developing any indicators they wish to use themselves or through their agent, QTAC.

Decisions about university selection — including decisions about course prerequisites, the evidence used in admission decisions and how that evidence is combined or weighted — are properly the responsibility of the universities. Universities remain free to use a range of evidence in selecting students for their courses. This evidence might include — but is not limited to — Year 12 results provided by QCAA (subject-specific and/or cross-curriculum), orders of merit based on overall achievement in senior studies and/or achievement in specific fields of study, special tests (such as tests of general ability), course-specific university entrance tests, interviews, portfolios, viva voce, lotteries, and the application of prerequisites for high-demand courses.

Implications of Principle 2
- Universities, as is their right, should continue to be responsible for deciding how their future students are selected, including by managing fair competition, where necessary, for high-demand courses.
- If universities choose to combine available evidence in some way, such as aggregating, scaling or weighting, then those processes are properly the responsibility of the universities themselves, not QCAA and the school sector.

Recommendations for separating certification and selection
8. QCAA should be responsible for the certification of student attainment at the end of Year 12 based on valid and reliable assessments but not for the calculation of rankings or other indicators that the universities might require.
9. Universities, through QTAC, should be responsible for comparing and ranking applicants from Year 12 to courses and for undertaking any associated scaling processes or other computations.

Principle 3 — Transparency, fairness

Processes for assessing student attainment in the secondary school and for selecting students for admission to universities should be as transparent as possible to students, parents and schools. Transparency is essential to fairness in assessment and selection processes.

Implications of Principle 3
- In addition to understanding how their achievements will be assessed and the criteria used to evaluate the quality of their work and performances, students should understand how their assessment results are combined to produce an overall result in each subject.
- Universities should make as transparent as possible the evidence to be used in course admission decisions, including processes for the selection of Year 12 completers who are ineligible for a tertiary entrance rank and for discriminating between eligible students when other measures have been exhausted, the use of bonus points, and offering places in advance based on school evidence or recommendation.
Consequential recommendations, in summary

10. Government should make legislative changes to divest QCAA of responsibilities relating to tertiary selection — Subject Achievement Indicators, QCS testing, QCS scaling parameters, generation of OPs and FPs, and other ad hoc measures such as QCS percentiles and an ATAR.

11. Universities should enhance technical capacity within QTAC to undertake any new scaling procedures for producing rank orders or deriving any other indicators that universities require.

12. Universities should review their admissions processes and consider options for comparing and selecting students. The review should consider the appropriateness of constructing a single rank order of Year 12 completers regardless of the course or institution to which they are applying, and options, apart from ATAR, for ranking course applicants.

Note for readers of the full report

There is no one-to-one relationship between the recommendations embedded in the discussion above and the formal list of 23 in the review report. Also, recommendations that appear in the review report on implementation, communication and governance are not discussed in this paper.

What was and what could be

Features of the proposed redesigned system juxtaposed with features of the existing system are in the diagram attached to this paper. Figure 1 illustrates what our proposed redesigned system would mean in practice if it were to be implemented: new Subject Results, new assessment package, new school assessment, new external assessment, new marking schemes for school assessments, new moderation model for school assessments, new certification, new tertiary entrance procedures, and new responsibilities ... in new times.

References


Senior assessment for students in Years 11 & 12

- School-based assessment retained and revitalised
- Moderation model revamped
- External assessment introduced
- Subject Results produced as standalone indicators of attainment

Tertiary entrance for Year 12 completers

- Universities and tertiary providers decide on method for using Subject Results to select Year 12 completers for entry

What is to be

- Prescribed and endorsed assessment types, conditions and marking schemes for three school assessments in each subject
- New marking schemes (criteria-based) with two-stage process for marking school assessments
- One external assessment in each subject to contribute 50 per cent to the Subject Result
- Results confirmed following each school assessment
- Subject Results produced by adding marks from three school assessments and one external assessment
- Subject Results reported from 1 to 60 (maximum)

What is no longer

- Levels of Achievement (from Very Limited Achievement to Very High Achievement)
- Grades (A–E) for each criterion in each assessment
- Overall grade (A–E) for each assessment criterion
- Standards matrices in each subject
- Consensus moderation using review panels
- Folios of student work for verification
- Queensland Core Skills testing for students
- Subject Achievement Indicators from teachers and schools
- Scaling to Queensland Core Skills Test group parameters
- Calculation of the Overall Position and Field Positions
- Calculation of an ATAR using Overall Achievement Indicators

What is to be

- Subject Results used as the basis for selecting students along with other criteria set by the universities and other providers
- Separation of responsibilities for senior certification and tertiary selection:
  - Queensland Curriculum and Assessment Authority is responsible for the certification of student attainment of Year 12 completers
  - Queensland Tertiary Admissions Centre is responsible for comparing and ranking applicants, and any scaling processes that might be necessary

What is no longer

- Overall Position
- Field Positions
- Queensland Core Skills Test grades
- Queensland Core Skills Test percentiles
- QTAC Selection Rank for OP-ineligible students

Figure 1 Redesigning the secondary–tertiary interface: Proposed new architecture (Matters & Masters, 2014)
Assessing general capabilities

Julian Fraillon
Australian Council for Educational Research

Julian Fraillon is the Research Director, Assessment and Reporting: Mathematics and Science within the ACER Assessment and Reporting Research Program. He has directed projects including the Australian National Assessment Program — Civics and Citizenship and the National Assessment Program — ICT Literacy, and coordinated the Asian Regional Module for the International Association for the Evaluation of Educational Achievement (IEA) International Civics and Citizenship Education Study 2009 (ICCS). Julian has worked on a great number of assessment programs for both academic and non-academic outcomes of schooling across Australia at local, state and territory and national levels. As well as his international work on the ICCS, Julian is an ongoing assessment consultant to the Hong Kong Education Bureau and has completed other assessment and monitoring consultancy work in East Timor and Chile and Australia. He is an Associate Editor for the Australian Journal of Education.

Juliette Mendelovits
Australian Council for Educational Research

Juliette Mendelovits is a Research Director at the Australian Council for Educational Research. She heads the Assessment and Reporting: Humanities and Social Sciences research program. She specialises in framework and test development for national and international large-scale learning assessments, with a particular focus on reading. Recent work includes a review of a range of widely-used reading assessments relevant to the Asia and Pacific regions; an evaluation of the reading tools and processes of citizen-led assessments that are conducted in India and Africa; and contribution to the development of the reading learning metric for the Learning Metrics Partnership. Her clients have included the Organisation for Economic Co-operation and Development, Educational Testing Service, Australian state and commonwealth government agencies and a number of national government ministries. She is an Associate Editor for the Australian Journal of Education.

Abstract

There is growing interest in general capabilities and cross-curricular learning outcomes such as literacy in information and communication technologies, creative thinking and collaborative and individual problem-solving. As the expectation for such competencies to be taught in schools has increased, so has the need for teachers and schools to validly and reliably assess student learning in those areas, and to report on them in ways that inform future teaching and learning. In this presentation we examine the challenges of assessing and reporting on student learning and learning growth in general capabilities and cross-curricular learning areas. We present approaches used in research to address some of these challenges and reflect on how these can be applied in the classroom.
Poster presentations
‘Assessment will play a central role in supporting implementation of new directions in science education internationally.’
— Professor James W Pellegrino
Midwives care for new life: Let’s care about new ways of recording our competencies
Lisa Charmer, Dr Denise Winkler and Dr Sally de-Vitry Smith
Charles Sturt University, NSW
An electronic portfolio (eportfolio) is in development to replace the paper-based workbook and diary used to record the clinical experiences of student midwives. These experiences capture the competencies required for registration as a midwife. This project is funded by a seed grant. The pilot of the portfolio will be the focus.
Charles Sturt University midwifery students, facilitators and academics will participate in a pilot of the eportfolio at a local regional hospital. Training will be provided to all participants with the use of mobile devices. Students will be asked to record their experiences using the eportfolio and evaluate the process. Focus groups will be held for participants to share examples of their experiences using three different formats. A follow-up electronic survey will be used to capture the thoughts and opinions emerging after the focus groups. Data will be analysed to refine the process prior to implementation across 60 clinical sites.

Transforming novice writers to discourse experts
Dr Tiffany Ip
The University of Hong Kong
In response to the curriculum change, an English-in-the-discipline course — ‘Dissertation Writing in the Social Sciences’ — is newly offered for third-year university students. Guided discovery and process writing approaches are employed in this new course development. This study sets out to investigate students’ perceived effectiveness at the tasks that were developed to tackle their writing problems. The first part of the study involves an examination of students’ perceived and actual writing difficulties. A questionnaire survey is carried out to understand students’ perceptions of the major writing difficulties. Student writing scripts are also analysed. The second part of the study involves an exploration of students’ views on the course effectiveness — whether the course is able to address their writing needs and difficulties. Interviews are administered to collect students’ opinions (together with teachers’ opinions as supplementary information) about the course content, learning activities and assessments.

Catch Up Literacy in Australia
Tracy Riley
Catch up Ltd, Victoria
This poster presents an overview about how and why Catch Up Literacy was piloted in Australia. It covers the development of Catch Up Literacy since the initial pilot, and includes examples of trainee feedback, case studies and success stories. Finally, the poster outlines what Catch Up wants to achieve in Australia in the future. The poster will contain a mixture of text, graphs and photographs to show students’ progress. There will also be quotes from teachers, parents and children.

Evaluating serious games with the Quantitative Evaluation Framework
Paul Escudeiro, Nuno Escudeiro, Marcelo Norberto, Jorge Lopes
ISEP – Instituto Superior de Engenharia do Porto, Portugal
This poster presents the overall evaluation of the Quantitative Evaluation Framework approach applied in an operational teaching environment. This environment includes the development of a serious game, supported by a web platform and extended to mobile platforms, which is being supervised by the research group GILT (Games, Interaction and Learning Technologies). The serious game, named Virtual Sign, aims to make the process of learning sign language easy and enjoyable. It is a funded quasi-experimental educational software project, which is being developed under the frame of a quality evaluation environment that measures system quality throughout its development life cycle. The quality evaluation process started with a careful planning phase. It has included the purpose of the evaluation, the timing of the evaluation and who should be conducting the evaluation process. Moderating the development of Virtual Sign with the Quantitative Evaluation Framework assures the quality of the final product.

Using a mixed approach to measure and assess the academic impact of an innovative Year 12 study program at Brighton Grammar
Ray Swann and Dan Belluz
Brighton Grammar School, Victoria
A central challenge in innovative curriculum piloting is how to capture the impact of the program simply and accurately. At Brighton Grammar, a new study program for Year 12 was piloted in 2014 and this poster illustrates how both a simple qualitative and quantitative framework was used. Specifically, each participant’s achieved versus predicted scores were plotted using VASS (Vitorian Assessment Software System). VASS includes a number of reports that can be used that control for the effect of both gender and ability. From a report based on the predicted versus actual performance for each subject included in the pilot, a subject differential was created. As this number did not include the teacher impact, a separate VASS report was used to control for this variable. The net loss or gain of the program was then calculated. As the effect of co-variance was not accounted for, an average across the piloted/non-piloted subjects was then used to measure the final impact, along with the results of the qualitative instrument. The final result was then used to analyse the impact of the pilot in a school setting.
Conference program
‘Authentic demonstrations of capability involve real tasks where students choose and justify the best course of action, actively employing their new knowledge and skills.’

— Dr Rosemary Hipkins
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.00–3.00</td>
<td>Registration</td>
</tr>
<tr>
<td>3.00–5.00</td>
<td><strong>Workshops</strong></td>
</tr>
<tr>
<td></td>
<td><strong>SESSION A</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Creating stealth assessments</strong></td>
</tr>
<tr>
<td></td>
<td>Prof Val Shute, Florida State University, USA; and Dr Michael Timms, ACER</td>
</tr>
<tr>
<td></td>
<td>Please bring your own device.</td>
</tr>
<tr>
<td></td>
<td><strong>SESSION B</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Assessing young children's literacy and mathematics understandings</strong></td>
</tr>
<tr>
<td></td>
<td>Prof Joanne Mulligan, Macquarie University; Marion Meiers, ACER; Maurice Walker and Prue Anderson, ACER</td>
</tr>
<tr>
<td></td>
<td><strong>SESSION C</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Improving assessment for Indigenous students</strong></td>
</tr>
<tr>
<td></td>
<td>Alison Quin, Queensland University of Technology</td>
</tr>
<tr>
<td></td>
<td><strong>SESSION D</strong></td>
</tr>
<tr>
<td></td>
<td><strong>New measures for an old friend: A learning progression for ICT literacy</strong></td>
</tr>
<tr>
<td></td>
<td>Prof Mark Wilson, The University of California, Berkeley and The University of Melbourne</td>
</tr>
<tr>
<td>5.30–7.00</td>
<td><strong>Networking drinks, Crown, Southbank, Melbourne</strong></td>
</tr>
</tbody>
</table>
| Time         | Session/Activity                                                                 
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>8.00–9.00</td>
<td>Registration</td>
</tr>
<tr>
<td>9.00–9.30</td>
<td>Welcome and Conference opening</td>
</tr>
</tbody>
</table>
| 9.30–10.45  | Keynote 1                                                                         
|             | Learning assessments: Designing the future                                       |
|             | Prof Geoff Masters AO, CEO, ACER                                                  |
|             | with pre-recorded contributions from Prof Dylan Wiliam, Institute of Education, UK|
| 10.45       | Morning tea                                                                       |
| 11.15–12.30 | Concurrent session Block 1                                                        |
|             | **SESSION E**                                                                     |
|             | Reflecting on teacher research on assessment: Challenges and motivations for the future |
|             | Prof Marie Brennan, Victoria University                                          |
|             | **SESSION F**                                                                     |
|             | Should generic curriculum capabilities be assessed?                               |
|             | Dr Rosemary Hipkins, New Zealand Council for Educational Research                |
| 12.30–1.30  | Lunch                                                                             |
| 1.30–2.45   | Keynote 2                                                                         
<p>|             | Assessment to action: New thinking from India                                     |
|             | Dr Rukmini Banerji, Pratham and the ASER Centre, India                           |
| 2.45–4.00   | Concurrent session Block 2                                                        |
|             | <strong>SESSION I</strong>                                                                     |
|             | Assessment standards, 'intentional alignment', and dialogic inquiry               |
|             | Prof Claire Wyatt-Smith, Australian Catholic University                           |
|             | <strong>SESSION J</strong>                                                                     |
|             | Assessment in interactive learning environments                                   |
|             | Dr Michael Timms, ACER; and Dr Jason Lodge, The University of Melbourne           |
| 6.30 for 7.00pm start | Conference dinner                                                                 |
|             | <strong>SESSION G</strong>                                                                     |
|             | Collaborative problem-solving: Assessment and reporting                           |
|             | Prof Patrick Griffin and Assoc Prof Esther Care, The University of Melbourne      |
|             | <strong>SESSION H</strong>                                                                     |
|             | PISA: Behind the headlines and past the rankings                                 |
|             | Dr Sue Thomson, ACER; and Chris Wardlaw PSM, Victorian Curriculum and Assessment Authority|
|             | <strong>SESSION I</strong>                                                                     |
|             | Assessment standards, 'intentional alignment', and dialogic inquiry               |
|             | Prof Claire Wyatt-Smith, Australian Catholic University                           |
|             | <strong>SESSION J</strong>                                                                     |
|             | Assessment in interactive learning environments                                   |
|             | Dr Michael Timms, ACER; and Dr Jason Lodge, The University of Melbourne           |
|             | <strong>SESSION K</strong>                                                                     |
|             | Conversation                                                                      |
|             | Dr Rukmini Banerji, Pratham and the ASER Centre, India                           |
|             | <strong>SESSION L</strong>                                                                     |
|             | Measuring what matters: Challenges and opportunities in assessing science proficiency |
|             | Prof James W Pellegrino, University of Illinois at Chicago, USA                    |</p>
<table>
<thead>
<tr>
<th>Time</th>
<th>Session M</th>
<th>Session N</th>
<th>Session O</th>
<th>Session P</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.00–10.15</td>
<td>Hacking assessment</td>
<td>The move to NAPLAN online: The advantages and the road ahead</td>
<td>Translating rich learning assessments into certified results and university selection devices</td>
<td>Assessing general capabilities</td>
</tr>
<tr>
<td></td>
<td>Assoc Prof Phillip Dawson, Deakin University</td>
<td>Dr Stanley Rabinowitz, Australian Curriculum, Assessment and Reporting Authority; with discussant Prof Barry McGaw, McGaw Group</td>
<td>Dr Gabrielle Matters, ACER</td>
<td>Julian Fraillon and Juliette Mendelovits, ACER</td>
</tr>
<tr>
<td>10.15–10.45</td>
<td>Morning tea</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.45–12.00</td>
<td>Concurrent session Block 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SESSION M</td>
<td>SESSION N</td>
<td>SESSION O</td>
<td>SESSION P</td>
</tr>
<tr>
<td></td>
<td>Hacking assessment</td>
<td>The move to NAPLAN online: The advantages and the road ahead</td>
<td>Translating rich learning assessments into certified results and university selection devices</td>
<td>Assessing general capabilities</td>
</tr>
<tr>
<td></td>
<td>Assoc Prof Phillip Dawson, Deakin University</td>
<td>Dr Stanley Rabinowitz, Australian Curriculum, Assessment and Reporting Authority; with discussant Prof Barry McGaw, McGaw Group</td>
<td>Dr Gabrielle Matters, ACER</td>
<td>Julian Fraillon and Juliette Mendelovits, ACER</td>
</tr>
<tr>
<td>12.00–1.00</td>
<td>Lunch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.00–2.30</td>
<td>Panel Session</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Designing the future for assessment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Panel speakers to include Prof Marie Brennan, Victoria University; Prof James W Pellegrino, University of Illinois at Chicago, USA; and Prof Val Shute, Florida State University, USA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conference close</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prof Geoff Masters AO, CEO, ACER</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Crown floor plan
Excellence in Professional Practice Conference 2016

Thursday 19 May & Friday 20 May
Bayview Eden, Melbourne

Contribute your knowledge and expertise to a conference presented by practitioners, for practitioners.

Call for proposals opens Saturday, 1 August 2015
Closes Friday, 4 December 2015

Registration: www.acer.edu.au/eppc

Enquiries: Margaret Taylor
T: 03 9277 5403   E: margaret.taylor@acer.edu.au
Rolling Summit on Assessment Reform and Innovation

ACER wants to hear from you. We are holding conversations around reform and innovation of student assessments, with face-to-face and online events as well as the Excellence in Professional Practice Conference and the Research Conference.

We call this a Rolling Summit because it’s growing in energy as it moves across the country. We aim to collect hundreds of examples of practice that supports student growth and success experiences. We also want to hear how you’ve overcome any barriers to change. We’ll share these at events and on our website throughout the year.

Principals, curriculum leaders and teachers: please let us know what you’re doing and how you’re thinking about assessment by going to https://www.surveymonkey.com/s/ACERassessment

If you would like to know more about the work of the Centre for Assessment Reform and Innovation, visit http://www.acer.edu.au/cari

The Centre for Science of Learning

The Centre for Science of Learning @ ACER is at the forefront of a new transdisciplinary field that we believe has the potential to improve teaching and learning.

Research in a number of disciplines is providing deeper insight into the nature of human learning. Advances in neuroscience, cognitive psychology, social and behavioural sciences and education are adding to our understanding of fundamental learning processes and of the conditions that lead to successful learning.

Our work in the Centre for Science of Learning @ ACER allows us to bring methods from those various disciplines together to gain better understandings of the important role of emotions, learning environments and effective pedagogical practices in education, and these insights will have direct implications for teaching and learning in educational settings.

Applying neuroscience, cognitive psychology and education, the Centre for Science of Learning @ ACER is developing evidence based strategies for learning, evaluating existing strategies, and creating a powerful narrative about the role of aspects of the brain in learning.

The Centre for Science of Learning @ ACER is a major partner in the Science of Learning Research Centre (SLRC), which is a Special Research Initiative of the Australian Research Council. In the Centre for Science of Learning @ ACER, researchers in education, neuroscience and cognitive psychology are working together with teachers to understand the learning process.

For more information, visit http://www.acer.edu.au/csl
Australian Council for Educational Research

On completion of this four-unit course participants will:
- Understand the theories and research evidence underpinning the purposes and principles of assessment and feedback in the teaching and learning cycle.
- Understand the uses of moderation and triangulation of data sources to inform judgements.
- Critically evaluate assessment in relation to defined frameworks.
- Critically evaluate a range of assessment methods, and use appropriate criteria to select and judge evidence.
- Build students’ capacity for self assessment and peer assessment.
- Use appropriate criteria to make unbiased judgements of student achievement based on evidence.
- Use assessment evidence to inform and improve current practice, identify next steps for students and identify professional development needs.

Graduate Certificate of Education
(Assessment of Student Learning)

Are you...
- a primary or secondary school teacher, coach or leader in education
- wishing to improve student learning
- seeking a recognised qualification to progress your career?

If you have a Bachelor’s degree (or higher) and access to a school setting for project work, consider the possibility of study with ACER.

For further details, including fees and dates, please check the website at http://courses.acer.edu.au or email courses@acer.edu.au
Dr Elizabeth Hartnell-Young
Director, ACER Institute

Australian Council for Educational Research
On completion of this four-unit course participants will:

Ø Understand the theories and research evidence underpinning the purposes and principles of assessment and feedback in the teaching and learning cycle.

Ø Understand the uses of moderation and triangulation of data sources to inform judgements.

Ø Critically evaluate assessment in relation to defined frameworks.

Ø Critically evaluate a range of assessment methods, and use appropriate criteria to select and judge evidence.

Ø Build students' capacity for self assessment and peer assessment.

Ø Use appropriate criteria to make unbiased judgements of student achievement based on evidence.

Ø Use assessment evidence to inform and improve current practice, identify next steps for students and identify professional development needs.

Graduate Certificate of Education
(Assessment of Student Learning)

Are you...

Ø a primary or secondary school teacher, coach or leader in education

Ø wishing to improve student learning

Ø seeking a recognised qualification to progress your career?

If you have a Bachelor's degree (or higher) and access to a school setting for project work, consider the possibility of study with ACER.

For further details, including fees and dates, please check the website at http://courses.acer.edu.au or email courses@acer.edu.au

Dr Elizabeth Hartnell-Young
Director, ACER Institute

Graduate study with ACER
Improving STEM Learning

What will it take?

7–9 August 2016
Brisbane Convention and Exhibition Centre

Research Conference 2016

Australian Council for Educational Research