# Student assessment regimes

## Getting the balance right for Australia

DRAFT discussion paper

**June 2009**



Queensland Government
Queensland Studies Authority

Queensland Studies Authority

# Contents

# 1. Purpose

Australia is entering an era of unprecedented national endeavour in education, with the recently established Australian Curriculum, Assessment and Reporting Authority (ACARA) at the helm.

It is well established that assessment and reporting have immense backwash effects on curriculum, influencing what is taught to students and how it is taught, and that these effects are particularly pronounced when teachers and schools operate in a regime of public accountability for students' results.

The intentions of assessment and reporting regimes are to ensure improved learning outcomes for students and provide valid and reliable data for policymakers.

There is compelling evidence from other countries, however, that when accountability for educational outcomes is measured solely by results in national full-cohort tests, the negative effects on teaching and student learning outweigh the positive intentions, and furthermore, that the data from such tests cannot be used by policymakers in meaningful ways.

This paper examines evidence from three countries — the United Kingdom, the United States of America and Australia — about the effects that assessment regimes have on student learning outcomes, how to avoid unbalanced assessment regimes that have unintended negative effects on those outcomes, and how national tests and teacher judgment can be used effectively to improve teaching and learning and provide meaningful data to policymakers.

The paper proposes principles for a balanced national assessment regime for Australia, in which accountability for student learning is shared by teachers, schools, schooling authorities and government agencies.

# 2. Background

Australian education ministers have made the following commitment to assessment of student achievement.

> Assessment of student progress will be rigorous and comprehensive. It needs to reflect the curriculum, and draw on a combination of the professional judgment of teachers and testing, including national testing.
>
> To ensure that student achievement is measured in meaningful ways, State, Territory and Commonwealth governments will work with all school sectors to develop and enhance national and school-level assessment that focuses on:
>
> - assessment for learning — enabling teachers to use information about student progress to inform their teaching
> - assessment as learning — enabling students to reflect on and monitor their own progress to inform their future learning goals
> - assessment of learning — assisting teachers to use evidence of student learning to assess student achievement against goals and standards
>
> (MCEETYA, 2008)

The professional judgment of teachers is integral to assessment and has been the ruling paradigm in Australia in the early and middle years of schooling. In most Australian jurisdictions, students sit subject-based external examinations at the end of Year 12. In the Australian Capital Territory (ACT) and Queensland, achievement standards are externally moderated in the final years of schooling.

National standardised tests are a relatively new phenomenon in Australia.

Students in Australian schools participate in sample triennial tests in Information Communication Literacy in Years 6 and 10, Science Literacy in Year 6, and Civics and Citizenship in Years 6 and 10. They also participate in full-cohort annual tests in Literacy and Numeracy in Years 3, 5, 7 and 9.

Australian students also participate in international tests. The Program for International Student Assessment (PISA) is conducted every three years with a sample of 15-year-old students and the Trends in International Mathematics and Science Study (TIMSS) is conducted every four years with a sample of students in Years 4 and 6.

Aggregated results on these tests are reported at a jurisdictional (State and Territory) level. However, the draft-for-consultation *MCEETYA Action Plan 2009–2012* (MCEETYA, 2008) flags intentions to measure school performance and track individual students' performance over time. The only student results available to measure school performance are the results from the full-cohort annual tests called the National Assessment Program — Literacy and Numeracy (NAPLAN).

# 3. Lessons from the United Kingdom and the United States of America

**Full-cohort standardised testing in an accountability environment has unintended negative effects on education.**

In general, the intentions of full-cohort standardised testing programs are to improve teaching and learning and provide valid and reliable data to policymakers. However, the research evidence from the United States of America (USA) and the United Kingdom (UK), where such programs have been in operation for many years, is that these tests have profoundly negative effects on teaching and learning, and the data they provide are not capable of informing policy decisions in meaningful ways.

## 3.1 Impact on learners

Full-cohort testing often reduces the self-esteem of lower-achieving students and makes it harder to convince lower-achieving students that they can succeed in other tasks (Assessment Reform Group 2002; Griffin & Heidorn 1996; Harlen & Deakin Crick 2002). Consistent evidence of poor performance can result in long-lasting loss of confidence (Stiggins 2009). Furthermore, repeated practice tests reinforce the low self-image of the lower-achieving students, ensuring that the gap between their achievements and those of higher-achieving students widens (Assessment Reform Group 2006; Harlen & Deakin Crick 2002).

## 3.2 Impact on teaching

Full-cohort tests encourage methods of teaching that promote shallow and superficial learning rather than deep conceptual understanding and the kinds of complex knowledge and skills needed in modern, information-based societies (Assessment Reform Group 2006; Shepard 2000, 2008; Pellegrino, Chudowsky & Glaser 2001). Teachers adopt transmission styles of teaching and highly structured activities (Harlen & Deakin Crick 2002). In order to secure higher test results for their students, teachers “teach to the test” and train students to pass the test, with consequent narrowing of the curriculum to what is tested and what can be tested (Harlen & Deakin Crick 2002; Herman, Baker & Linn 2006; Jennings & Rentner 2006; Koretz 1988; Linn 1998, 2000; Popham 2001; Shepard, 2008).

Full-cohort tests rarely provide information that teachers can use to improve their teaching and student learning (Pellegrino, Chudowsky & Glaser 2001) especially when the major use of test results is to determine whether students have met a minimal standard or benchmark (Herman, Baker & Linn 2006).

## 3.3 Misinterpretation of test results

Although the proponents of full-cohort tests claim that these tests boost standards of achievement, there is no firm evidence to support this claim (Assessment Reform Group 2006). Rather, the claims are based on misinterpretations of test results (Stiggins 2007).

It is often assumed that increased test scores over time indicate that students’ learning has increased. However, it has been convincingly demonstrated that these increases are often due to a combination of teachers “teaching to the tests” and students becoming familiar with the tests (Assessment Reform Group 2006; Koretz 1988; Linn 2000, 2001; Shepard 2000; Wiliam 2008b). In reality, students often do *not* have the knowledge and skills that the test results supposedly indicate they have (Harlen & Deakin Crick 2002; Koretz, et al. 1991).

While full-cohort standardised tests are usually so narrowly focused that they are capable of providing little information about student achievement, results are often used as if they provide a great deal of information about how well students have learned and how well schools have taught their students (Harlen & Deakin Crick 2002). Clearly, this is a misuse of information.

In addition, there are usually large measurement errors in test results, yet they are often used as if the information is quite precise (Assessment Reform Group 2006; Black & Wiliam 1998; Harlen & Deakin Crick 2002; Koretz 1988; Linn 1998, 2001; Popham 2003; Wiliam 2008b). As a result of such use, students are frequently given the wrong levels, and unfair and incorrect decisions are made about some pupils (Assessment Reform Group 2006). When used for school accountability purposes, the probability of misclassification, especially for small schools, is disturbingly large (Linn 1998).

In fact, the nature of full-cohort tests, with their imprecise and single point-in-time results, renders them incapable of providing useful and valid information to policymakers. The high financial and time burdens at national and school levels are not justified by the limited value of the information gained (Assessment Reform Group 2006; Pellegrino, Chudowsky & Glaser 2001).

## 3.4 School accountability for test performance

In school accountability regimes, results of tests are frequently used to judge the quality of teaching in schools. However, studies indicate that very little of the variability in test scores can be attributed to the school (Popham 2003; Stake 2007; Wiliam 2008b).

When schools are rewarded or sanctioned for student results on full-cohort standardised tests, there are enormous pressures on the schools to adopt practices — however nefarious — to elevate test scores. Known practices include, but are not limited to, excluding low-achieving students from the school, classifying low-achieving students as having disabilities so that they do not have to sit the tests, holding students in a lower grade for a year and then having them jump the grade in which the tests are held, and encouraging students who feel they are going to fail, yet again, to voluntarily drop out of school (Heilig & Darling-Hammond 2008). Teachers and school administrators have also blatantly cheated, both by teaching actual test items just before a test and giving students actual test items (Koretz 1988).

Even the National Assessment of Educational Progress (NAEP) sample testing program in the USA seems to have accrued some of the undesirable side effects of full-cohort testing once the results began to be used in an environment of accountability and state-by-state comparisons. As results were used to measure schools' performance, teachers began to "teach to the test", resulting in a narrowing of the curriculum and diminished teaching oriented to student diversity. There is recognition that NAEP scores do not adequately represent the quality of schooling, which is complex and therefore not able to be reduced to scaled scores on a test (Stake 2007).

## 3.5 The USA and UK in summary

The use of accountability-oriented standardised testing programs in the USA has failed to deliver positive educational outcomes, as summarised in the following statement by Robert Linn:

> As someone who has spent his entire career doing research, writing and thinking about educational testing and assessment issues, I would like to conclude by summarizing a compelling case showing the major uses of tests for student and school accountability during the last 50 years have improved education and student learning in dynamic ways. Unfortunately, that is not my conclusion. Instead, I am led to conclude that in most cases the instruments and technology have not been up to the demands that have been placed on them by high-stakes accountability. Assessment systems that are useful monitors lose much of their dependability and credibility for that purpose when high stakes are attached to them. The unintended negative effects of the high-stakes accountability uses often outweigh the intended positive effects.
>
> (Linn 2000, p. 14)

Lorrie Shepard provides further evidence:

> … external accountability testing leads to the de-skilling and de-professionalization of teachers, even — in my own state recently — to the denigration of teaching. High-stakes accountability teaches students that effort in school should be in response to externally administered rewards and punishment rather than the excitement of ideas. And accountability-testing mandates warn teachers to comply or get out (or move, if they can, to schools with higher scoring students).
>
> (Shepard 2000, p. 9)

> Given several decades of high-stakes, test-based accountability, it is conceivable that such programs are sometimes the *cause* of poor instruction and limited learning …
>
> (Shepard 2008, p. 43)

The overall negative impact on education of full-cohort testing and the accountability regime in the UK has been recorded in the House of Commons' *Select Committee on Children, Schools and Families Third Report* to the United Kingdom Parliament in May 2008. The negative impacts recorded in this report reinforce the findings of UK researchers over the last twenty years as discussed previously. The negative impacts recorded by the Select Committee are summarised below:

- teachers have been forced to "teach to the test", thereby narrowing the educational experiences and attainments by students
- schools have pursued test results at the expense of a rounded education for children
- it is possible to improve test scores through mechanisms such as "teaching to the test", narrowing the curriculum and concentrating effort and resources on borderline students
- these classroom practices have distorted the education of some students, leaving them unprepared for higher education and employment
- the improvement in test scores does not necessarily provide evidence of enhancement of underlying learning and understanding in pupils
- pupils may not retain or may not even possess in the first place, the skills which are supposedly evidenced by their test results
- many students have not received their entitlement to learning due to the demands of national full-cohort testing
- measurement of standards across the full curriculum is virtually impossible under a regime of full-cohort tests.

# 4. Ideas for a balanced national assessment program from the UK and the USA

The House of Commons' *Select Committee on Children, Schools and Families Third Report* to The United Kingdom Parliament in May 2008 also made a number of recommendations for improving the national assessment regime. These are summarised below:

- to avoid negative consequences of using assessment results for accountability purposes use both sample test results and the results of teacher assessment
- the purpose of national monitoring of the education system is best served by sample testing to measure standards over time
- in the interests of public confidence such sample testing should be carried out by a body at arms length from the government
- teacher assessment should form a significant part of a national assessment regime
- assessment for learning should be supported by enhanced professional development for teachers.

As a consequence of this report, in 2008 the UK Schools Secretary, Ed Balls, announced that he was ending the requirement that schools run national tests for 14-year-olds, and that the national tests would be replaced by improved assessment by teachers (The United Kingdom Department for Children 2008).

The Assessment Reform Group in the UK has, for many years, documented how teachers' assessment capabilities can be enhanced, especially when the results are to be reported or used for accountability purposes.

> The following five points were brought out in consulting those with experience of implementing summative assessment with teachers in the UK and in Australia and the United States (Queensland and California) …
>
> 1. Teachers should have clear criteria describing levels of progress in various aspects of achievement, and, ideally, they should help to develop these criteria …
> 2. Professional development is needed so that teachers follow procedures that ensure dependability …
> 3. A system of moderation of teachers' judgments through professional collaboration benefits teaching and learning as well as assessment. Moderation that affects the planning and implementation of assessment, and consequently teachers' understanding of learning goals and of the criteria indicating progress towards them, has more than a quality assurance function.
> 4. The provision of a bank of well-designed tasks, with marking criteria … [is] part of the evidence that teachers can use, if needed, to ensure that all intended goals are taken into account in their assessment …
> 5. … It is important that all involved have time to trial and evaluate new practices and to be clear about the procedures and the safeguards that are built in to protect dependability.
>
> (Assessment Reform Group 2006, p. 4)

The same group also advocated the following actions for national and local policymakers:

> – Recognise that the financial and time burdens at national and school levels of current summative assessment policies based on testing are not justified by the value of the information gained.
> – Replace national testing, where it exists, by a requirement for reporting moderated judgments of pupil performance, and divert some of the time and money saved into quality assurance that enhances teaching and learning …
> – Set up a system of sampling pupils' performance for national monitoring, thereby reducing the overall test burden whilst increasing the breadth and relevance of the evidence.
>
> (Assessment Reform Group 2006, p. 13)

The National Research Council in the USA (Pellegrino, Chudowsky & Glaser 2001) made a number of recommendations for policy and practice in that country. These included the following:

- Large-scale assessment tools and supporting instructional material should have clear learning goals.
- The assessment criteria should be available to all potential candidates and other concerned individuals.
- A variety of matrix sampling, curriculum-embedded tools should be use to cover the breadth of a domain of the curriculum.
- Instruction on how students learn and how learning can be assessed should be a major component of teacher pre-service and professional development programs.
- Important decisions about individuals should not be based on a single test score.
- The balance of mandates and resources should be shifted from an emphasis on external forms of assessment to an increased emphasis on classroom formative assessment.
- Efforts should be made to foster public understanding of basic principles of appropriate test interpretation and use.

# 5. A balanced national assessment system for Australia

Australian policymakers are urged not to repeat the mistakes of the UK and the USA. Australia can establish assessment policy that will support the Australian education ministers' commitments to a national assessment system that will:

- be comprehensive, rigorous and meaningful
- improve teaching and learning
- draw on the professional judgment of teachers and national testing.

Getting the right balance between teachers' professional judgment and national testing is crucial for national assessment systems to be comprehensive, rigorous and meaningful while, at the same time, improving teaching and learning.

## 5.1 National testing

The research evidence is unequivocal on the use of full-cohort standardised test results for accountability purposes — they *cannot* be used to measure the effectiveness of teaching in schools. Furthermore, practices in the UK and the USA of holding schools accountable for the test results of their students have resulted in impoverished educational outcomes for both countries, and counter-productive stress on students, teachers and school communities.

Sample tests are cheaper to administer than full-cohort tests, and they do not have the negative consequences of teachers being forced to teach to the test, thereby narrowing the curriculum and the range of student achievements.

Importantly, the results of sample tests cannot be used for the destructive practice of comparing one school's results with another school's results in the mistaken belief that the effects of teaching and subsequent student learning are being compared, and that this will somehow improve student learning.

The results from national sample tests can, however, provide data to inform policy decisions regarding the health of Australian educational systems, and where resources should be deployed to ensure better educational outcomes.

## 5.2 Accountability and teacher judgment

Schools and schooling systems are held accountable for teaching and assessing students effectively. So too are those who prepare pre-service teachers accountable for providing development of the basic skills needed to construct assessment instruments and make defensible judgments about student achievement.

Schooling systems need to be accountable for supporting teachers to assess their students' progress and report the results of their assessments. Schooling systems can support teachers by providing online central repositories of high-quality assessment instruments connected to the curriculum teachers are expected to teach, and by providing continued professional development of teachers to become expert in constructing high-quality assessment instruments and making defensible judgments about their students' achievements.

The United Kingdom Parliament has acted to move the UK towards assessment practices that strengthen the assessment capabilities and, consequently, the teaching capabilities of teachers in that country. Some states in the USA have moved to include teacher assessment in their state assessment regimes (Darling-Hammond & McCloskey 2008).

The evidence from the UK, the USA and Australian researchers (Forster 2001; Forster & Masters 2004; Klenowski & Adie 2008; Klenowski & Wyatt-Smith 2008; Masters & Forster 2000) indicates the direction for national testing in Australia and provides evidence of how teachers' professional judgments can be given credibility in an accountability regime.

If teacher assessment is to be afforded credibility by the community, there must be accountability mechanisms to ensure that the reported results of teacher assessment are robust and comparable from school to school.

In the UK, the Qualifications and Curriculum Authority piloted a range of moderation models in 2007 and 2008. The report of the pilot contends that moderation is an essential part of an assessment system to ensure comparable outcomes and improve teachers' assessment capabilities:

> Sound moderation is a prerequisite of any assessment system, aimed at ensuring that agreed standards are applied consistently by the individuals involved. In assessing pupils' progress (APP), moderation should ensure that teacher assessment outcomes are reliable and comparable between school and across local authorities …
>
> Taking part in a real moderation proved to be a highly effective way of learning how to improve the quality of their [teachers'] own assessments as well as how to confirm the assessments of others. All the available evidence showed a positive impact on the nature and range of evidence used to support assessment, the accuracy of assessments and the understanding of what characterises performance at a national curriculum level.
>
> (Qualifications and Curriculum Authority 2009b, pp. 3, 24)

At the same time as the pilot studies were taking place in the UK, researchers have been undertaking a study into the effects of moderation in Years 1–10 in Queensland under an Australian Research Council Linkage Project. Preliminary findings of this research support the crucial role of moderation to improve teachers' assessment in an era of accountability:

> Moderation too is intrinsic to efforts by the profession to realize judgments that are defensible, dependable and open to scrutiny. Moderation can no longer be considered an optional extra and requires system-level support especially if, as intended, the standards are linked to system-wide efforts to improve student learning.
>
> (Klenowski & Wyatt-Smith 2008, p. 1)
>
> The initial stage of the research reported in this paper suggests that the practice at the local level of social moderation has the potential to fulfill an important role as a process for aiding teachers in ascribing value to student work through the use of standards that help them understand curriculum year level requirements and student achievement within year levels and in doing so attend to system level accountability.
>
> (Klenowski & Aidie 2008, p. 2)

As acknowledged by Klenowski & Wyatt-Smith (2008), Queensland has nearly forty years' experience in developing and implementing a system of moderation in which teachers and schools are accountable for the assessment and reporting of student achievement in Years 11 and 12. The inter-marker reliability of this moderation system surpasses that of many external examination regimes (Jordan & McDonald 2008; Masters & McBryde 1994; Stanley & Tognolini 2008).

The Queensland Studies Authority (QSA), which regulates the Queensland system of externally moderated school-based assessment, has recently published a set of principles for moderation to occur effectively. These principles have their foundation in the seminal work of Sadler (1986), the refinements of Pitman, O'Brien & McCollow (1999*)* and Matters (2006), and forty years of experience in moderation in an accountability environment as well as the research evidence discussed here in the international literature.

> For social moderation to work effectively, the following are required:
>
> – syllabuses that clearly describe content and achievement standards
>
> – contextualised exemplar assessment instruments

- samples of student work annotated to explain how they represent different standards
- consensus through teacher discussions on the quality of the assessment instruments and the standards of student work
- professional development of teachers
- an organisational infrastructure encompassing the QSA and schools to ensure the above takes place.

(Queensland Studies Authority 2009, p. 3)

# 6. Recommendations for a balanced student assessment system for Australia

The Australian education ministers' commitment to a robust assessment system must be supported by policy informed by sound research evidence. Government agencies, the jurisdictions, schools and teachers are responsible, and accountable, for assessment.

The commitments are that assessment of student achievement will:

- be comprehensive, rigorous and meaningful
- improve teaching and learning
- draw on the professional judgment of teachers and national tests.

The research evidence from the UK, the USA and Australia points us in the following directions.

## 6.1 National tests

- National tests are best used to estimate the health of the Australian education system.
- Data from national tests must not be used to compare schools in the mistaken belief that such comparisons will result in improved teaching and student learning.
- Data from national tests must be used ethically to improve teaching and student learning.
- National tests should take the form of sample tests constructed to assess students' achievements in the national curriculum.
- The accountability for constructing, analysing and reporting the results of national tests should reside in a statutory body, at arms length from government, but accountable to government for the quality of the tests and ensuing data.
- Jurisdictions must use the results of national sample tests to provide support for teachers in those areas of the curriculum that have been identified by the test results as needing such support.

## 6.2 The professional judgment of teachers

- Teachers can make valid and reliable assessments when they know what it is they are to teach, the criteria they are to use in assessing students' achievements, and engage in professional discussions with other teachers about the standards evident in their students' work.
- Teachers must be accountable for their own participation in professional development activities, assessing their students' progress and moderation of assessment, so that reported results are fair to all students and comparable from school to school.
- Schools must be accountable for providing supporting infrastructure for teachers to participate in professional development and moderation activities.
- The statutory body accountable for the national tests should also be accountable for providing a bank of high-quality assessment instruments (available online) for teachers to use with their students as tools for assessment of, for and as learning.
- Universities must be accountable for providing beginning teachers with the basic knowledge and skills to construct assessment instruments and assess their students' achievements.
- Jurisdictions must be accountable for providing the support necessary for teachers to improve their teaching, and student learning, in those aspects of the curriculum in which analyses of national testing data indicate support is needed.

- Jurisdictions must be accountable for providing ongoing professional development for teachers that enables them to construct high-quality assessment instruments and make assessments of student achievement that both improve teaching and learning and report on the progress of their students.
- Jurisdictions must be accountable for providing supporting infrastructure for teachers to participate in the moderation of assessment.

# Bibliography


Assessment Reform Group 1999, *Assessment for Learning: Beyond the black box,* University of Cambridge.

—— 2002, *Testing, Motivation and Learning*, ARG, Cambridge.

—— 2006, *The Role of Teachers in the Assessment of Learning*, ARG, Cambridge.

—— 2008, *Changing Assessment Practice: process, principles and standards*, ARG, Cambridge.

Australian Association of Mathematics Teachers Inc. 2008, *Position paper on the practice of assessing mathematics learning*, accessed 16 Mar 2008, <www.aamt.edu.au>.

Australian Council for Educational Research 2007, "How not to measure student performance", *ACER eNews*, issue 54, Jun.

Black, P & Wiliam, D 1998, *Inside the Black Box: Raising standards through classroom assessment,* School of Education, King's College, London.

—— 2004, "Classroom assessment is not (necessarily) formative assessment (and vice-versa)" in M Wilson (ed.) *Towards coherence between classroom assessment and accountability: 103rd Yearbook of the National Society for the Study of Education,* University of Chicago Press, Illinois.

Cambridge Assessment 2008, *Alternative Approaches to National Assessment at KS1, KS2 and KS3,* Cambridge Assessment, Cambridge.

Cizek, GJ 2005, "High-stakes testing: contexts, characteristics, critiques, and consequences", in Richard P Phelps (ed.) *Defending Standardized Testing,* Lawrence Erlbaum Associates, Mahwah, New Jersey.

Cumming, J & Maxwell, GS 2004, "Assessment in Australian schools: Current practice and trends", *Assessment in Education,* vol. 11, no. 1, pp. 89–108.

Darling-Hammond, L 1988, "Accountability and teacher professionalism", *American Educator,* vol. 12, pp. 8–13.

Darling-Hammond, L & McCloskey, L 2008, "Assessment for learning around the world: What would it mean to be internationally competitive?", *Phi Delta Kappan,* Dec, pp. 263–272.

Delandshere, G 2002, "Assessment as inquiry", *Teachers College Record,* vol. 104, no. 7, pp. 1461–1484.

Forster, M 2001, *A policy makers guide to system wide assessment programs*, Australian Council for Educational Research, Melbourne.

Forster, M & Masters, G 2004, "Bridging the conceptual gap between classroom assessment and system accountability", in M Wilson (ed.) *Towards coherence between classroom assessment and accountability: 103rd Yearbook of the National Society for the Study of Education,* University of Chicago Press, Illinois.

Frederickson, J & White, B 2004, "Designing assessment for instruction and accountability: An application of validity theory to assessing scientific inquiry", in M Wilson (ed.) *Towards coherence between classroom assessment and accountability, 103rd Yearbook of the National Society for the Study of Education Part II,* Chicago, National Society for the Study of Education, pp. 74–104.

Griffin, BW & Heidorn, MH 1996, “An examination of the relationship between minimal competency test performance and dropping out of high school”, *Educational Evaluation and Policy Analysis*, vol. 18, pp. 243–52.

Hall, K & Harding, A 2002, “Level descriptions and teacher assessment in England: Towards a community of assessment practice”, *Educational Research*, vol. 44, pp. 1–5.

Hargreaves, DJ, Galton, MJ & Robinson, S 1996, “Teachers’ assessments of primary children’s classroom work in the creative arts”, *Educational Research*, no. 38, pp. 199–211.

Harlen, W, 2004 “A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes”, in *Research Evidence in Education Library*, EPPI-Centre, Social Science Research Unit, Institute of Education, London.

—— 2005, “Trusting teachers’ judgment: Research evidence of the reliability and validity of teachers’ assessment used for summative purposes”, *Research Papers in Education,* vol. 20, no. 3, pp. 245–70, Sep.

Harlen, W & Deaken Crick, R 2002, “A systematic review of the impact of summative assessment and tests on students’ motivation for learning (EPPI-Centre Review, version 1.1)”, in *Research Evidence in Education Library*, issue 1, EPPI-Centre, Social Science Research Unit, Institute of Education, London.

Heilig, JV & Darling-Hammond, L 2008, “Accountability Texas-style: The progress and learning of urban minority students in a high-stakes testing context”, *Educational Evaluation and Policy Analysis*, vol. 30, no. 2, pp. 75–110.

Herman, JL, Baker, EL & Linn, RL 2006, “Assessment for accountability and learning”, *CRESST LINE, Newsletter of the National Center for Research on Evaluation, Standards, and Student Testing*, Fall edition.

House of Commons 2008, *Select Committee on Children, Schools and Families Third Report*, United Kingdom Parliament, May.

Jennings, J & Rentner, DS 2006, “Ten big effects of the No Child Left Behind Act on public schools”, *Phi Delta Kappan*, vol. 88, no. 2.

Jordan, P & McDonald, J 2008, data presented in a symposium “How standards of student achievement work to support teacher judgement: The place of moderation”, Australian Association for Research in Education Conference, Brisbane, Dec.

Klenowski, V & Adie, L 2008, “Moderation as judgement practice: Reconciling system level accountability and local level practice”*, Curriculum Perspectives,* in press*.*

Klenowski, V & Wyatt-Smith, C 2008, “Standards-driven reform Years 1–10: Moderation an optional extra?”, paper presented at the Australian Association for Research in Education Conference, Brisbane, Dec.

Koretz, D 1988, “Arriving at Lake Wobegon: are standardised tests exaggerating achievement and distorting instruction?”, *American Educator,* vol. 12 (2), no. 8–15, pp. 46–52.

—— 2008, “Further steps towards the development of an accountability-oriented science of measurement”, in K Ryan & L Shepard, *The Future of Test-based Educational Accountability.* Routledge, New York.

Koretz, D, Linn, L, Dunbar, SB & Shepard, LA 1991, “The effects of high-stakes testing on achievement: preliminary findings about generalisation across tests”, paper presented at the annual meeting of the American Educational Research Association, Chicago.

Linn, RL 1998, *Assessments and accountability,* CSE Technical report 490, CRESST, University of Colorado, Boulder.

—— 2000 "Assessments and accountability", *Educational Researcher,* vol. 29, no. 2, pp. 4–16.

—— 2001, *The design and evaluation of educational assessment and accountability systems*, Center for the Study of Evaluation Technical Report 539, University of California, Los Angeles.

—— 2008, "Educational accountability systems", in K Ryan & L Shepard, *The Future of Test-based Educational Accountability.* Routledge, New York.

Madaus, GF 1985, "Test scores as administrative mechanisms in educational policy", *Phi Delta Kappan,* vol. 66, no. 9, pp. 611–17.

Marion, SF & Sheinker, A 1999, *Issues and Consequences for State-level Minimum Competency Testing Programs,* National Center on Educational Outcomes, Jan.

Masters, GN, Ainly, J, Rowley, G & Khoo, ST 2008, "*Reporting and Comparing School Performances*, 2008", paper prepared for the MCEETYA Expert Working Group to provide advice on national schools data collection and reporting for school evaluation, accountability and resource allocation, Dec.

Masters, GN & Forster, M 2000, *The Assessments We Need*, Australian Council for Educational Research.

Masters, GN & McBryde, B 1994, *An Investigation of the Comparability of Teachers' Assessments of Student Folios*, Tertiary Entrance Procedures Authority, Brisbane.

Matters, G 2006, *Assessment Approaches in Queensland Senior Science Syllabuses*, Queensland Studies Authority, Brisbane.

Ministerial Council on Education, Employment, Training and Youth Affairs 2008, *MCEETYA Action Plan 2009–2012: A companion document for the Melbourne Declaration on Educational Goals for Young Australians*, (Draft-for-consultation), MCEETYA, 5 Dec.

—— 2008, *Melbourne Declaration on Educational Goals for Young Australians* (Draft), MCEETYA, Dec.

Pelliegrino, J, Chudowsky, N & Glaser, R 2001, *Knowing What Students Know: The science and design of educational assessment.* National Academy Press, Washington, DC.

Pitman, JA, O'Brien, JE & McCollow, JE 1999, "High quality assessment: We are what we believe and do", paper presented at the IAEA Conference, Bled, Slovenia, May.

Popham, WJ 2001, "Teaching to the test", *Educational Leadership,* vol. 58, no. 6, pp. 16–20.

—— 2003, *Why Standardized Tests Don't Measure Educational Quality*, EBSCO Publishing.

Qualifications and Curriculum Authority 2009 a, *Assessing Pupils' Progress: Putting the learner at the heart of good assessment,* Qualifications and Curriculum Authority, London.

—— 2009 b, *Report on trial of models of moderation within assessing pupils' progress 2007/8*, in press, London.

Queensland Studies Authority, 2009, *P–12 Assessment Policy*, QSA, Brisbane.

Rowley, G & Ingvarson, L 2007, "How not to measure teacher performance", *ACER News Issue 54*, Australian Council for Educational Research.

Ryan, K & Shepard, L (ed.) 2008, *The Future of Test-based Educational Accountability*, Routledge, New York.

Sadler, R 1986, *Defining and Achieving Comparability of Assessments,* Queensland Board of Secondary School Studies, Brisbane.

Shepard, LA 2000, "The role of assessment in a learning culture", *Educational Researcher,* Feb.

—— 2008, "A brief history of accountability testing 1965–2007", in K Ryan & L Shepard, *The Future of Test-Based Educational Accountability,* Routledge, New York.

Stake, RE 2007, "NAEP, Report Cards and Education; A Review Essay", *Education Review*, vol. 10, no. 1.

Stanley, G & Tognolini, J 2008, "Performance with respect to standards in public examinations", paper presented at the Annual Conference of the International Association for Educational Assessment, Cambridge, 10 Sep.

Stiggins, R 2004, "New assessment beliefs for a new school mission", *Phi Delta Kappan*, Sep, pp. 22–27.

—— 2007, "Five assessment myths and their consequences", *Education Week*, Oct.

—— 2008, *Assessment Manifesto: A call for the development of balanced assessment systems*, ETS Assessment Training Institute, Portland.

—— 2009, "Assessment, student confidence, and school success*", Phi Delta Kappan,* Feb.

The United Kingdom Department for Children 2008, "Schools and Families", *Press Notice 2008/0229,* 14 Oct.

Volante, L 2005, "Accountability, student assessment, and the need for a comprehensive approach", *International Electronic Journal for Leadership in Learning,* vol. 9, no. 6.

Wiggins, G 1989, "A true test: Toward more authentic and equitable assessment", *Phi Delta Kappan*, May, pp. 703–13.

Wiliam, D 2000, "Reliability, validity and all that jazz", *Education,* vol. 29, no. 3, pp. 9–13.

Wiliam, D 2008 a, "International comparisons and sensitivity to instruction", *Assessment in Education: Principles, Policy and Practice,* vol. 15, no. 3, pp. 253–57.

—— 2008 b, "What do you know when you know the test results? The meanings of educational assessments", keynote address at the Annual Conference of the International Association for Educational Assessment, Cambridge, Sep.

Wilson, M 2004, "Assessment, accountability and the classroom: A community of judgement", in M Wilson (ed.) *Towards coherence between classroom assessment and accountability: 103rd Yearbook of the National Society for the Study of Education*, University of Chicago Press, Illinois.

**Queensland Studies Authority**
295 Ann Street, Brisbane
PO Box 307 Spring Hill
QLD 4004 Australia
**T** 07 3864 0299
**F** 07 3864 0401
**www.qsa.qld.edu.au**