A Systems Level Approach to Building Sustainable Assessment Cultures: Moderation, quality task design and dependability of judgment

Peta Colbert, Claire Wyatt-Smith & Val Klenowski

Abstract
This paper considers the conditions that are necessary at system and local levels for teacher assessment to be valid, reliable and rigorous. With sustainable assessment cultures as a goal, the paper examines how education systems can support local level efforts for quality learning and dependable teacher assessment. This is achieved through discussion of relevant research and consideration of a case study involving an evaluation of a cross-sectoral approach to promoting confidence in school-based assessment in Queensland, Australia. Building on the reported case study, essential characteristics for developing sustainable assessment cultures are presented, including: leadership in learning; alignment of curriculum, pedagogy and assessment; the design of quality assessment tasks and accompanying standards, and evidence-based judgment and moderation. Taken together, these elements constitute a new framework for building assessment capabilities and promoting quality assurance.

Introduction
This past decade has witnessed an increased interest in assessment in a learning culture (Popham, 2008; Stobart, 2009; Darling-Hammond, 2010) with quality teacher assessment understood as central to local and system efforts to improve student learning and outcomes (Murphy, 2009; Willis, 2010; Wilson, 2010). There are signs of growing research and policy interest in the matter of how teacher assessment can be an option for summative or reporting purposes, and the conditions under which this option can be adopted with confidence (Harlen, 2004; Daugherty, 2008; Mansell, James, & Assessment Reform Group, 2009; Maughan, 2009). Specifically, this paper considers a system level approach to building sustainable assessment cultures and explores how it can support local level efforts for quality learning, dependable teacher assessment and local level curriculum development.

The case is made that it is timely to go beyond centrally controlled and centrally devised high-stakes assessment to include teacher-generated, locally devised assessment tasks for reporting and accountability purposes. Centrally maintained
processes for a sustainable assessment culture that support and provide opportunities for teacher-devised assessment are presented. These findings are based, in part, on an evaluation conducted in 2008 of the trial of the School-devised Assessment Model of Queensland Comparable Assessment Tasks (QCATs) (Wyatt-Smith & Colbert, 2008) and supplemented with further related research on teacher assessment, with specific reference to teacher designed assessments with accompanying statements of assessment standards (Stanley, MacCann, Gardner, Reynolds, & Wild, 2009, Klenowski, 2007). Recommendations are made on the basis of the evaluation and related studies on moderation (Klenowski & Wyatt-Smith, 2010; Klenowski & Adie, 2009) that have informed our view that accountability checks and balances can serve to maintain the rigour and reliability of teachers’ assessment. The conceptualised model presented in the paper offers opportunities for the public to have confidence in teachers’ judgments and assists in the building of more sustainable assessment cultures.

This paper argues for intelligent accountability and offers a dual approach to assessment, in effect meeting the purposes of learning improvement and reporting. The model seeks to achieve this duality. Specifically, it foregrounds the necessary professionalism of teachers and supports them in devising locally generated assessments for the improvement and support of learning, reporting and accountability purposes. In presenting this argument, conditions for promoting confidence in school-based assessment are presented first, followed by discussion of the context in which the evaluation and research were conducted. The methodology for the evaluation and analysis of the school-devised model follows. Our findings include insights into the processes, checkpoints and resources needed for quality school-based assessment, inter- and intra-school moderation practices and content for ongoing teacher professional development.

**Conditions for promoting confidence in school-based assessment**

Three main premises inform this discussion. First, classroom teachers are in the ideal situation to monitor their students’ learning and provide informed judgments of student achievement. Second, the community and the teaching profession are not well served when standards are used primarily as the instrument of government reporting and where those same standards are not routinely taken up by teachers in how they design and implement classroom assessment and monitor student achievement. Third is the understanding that the provision of a standards-referenced
framework is a necessary, though not sufficient, condition for assessment by teachers to be a dependable option for summative purposes. Other conditions are also necessary, including exemplars of well-designed assessment tasks, the installation of moderation practices with in-built quality assurance checks, and ongoing support for teachers in the form of appropriate professional development.

The term *dependability* has been considered a useful construct in assessment research in that it attends to issues of assessment design and implementation, with a focus on both quality of assessment items and comparability of awarded grades. Dependability is viewed as a combination of reliability and validity and involves the extent to which reliability is optimised while preserving validity (Harlen, 2004). To preserve validity there must be ‘a clear definition of the domain being assessed, evidence that in the assessment process the intended skills and knowledge are used by the learners’ (Harlen, p. 25), and *rigour* of assessment design. The latter points to the need for what the authors of the evaluation referred to above refer to as *cognitive stretch: that is, assessment* that extends beyond content knowledge to higher order analytic capabilities, critical thinking or evaluative skills, creativity, and the demonstrated portability of knowledge from one context of use or application to others. Working from this position, a dependable process for assessing and judging student achievement requires a focus on quality assessment design processes that reflect domain objectives and that are able to produce evidence of deep thinking and intellectual rigour in student learning.

If school-based assessment is to have utility for students’ learning (formative) and reporting purposes (summative) at system level, a range of strategies aimed at supporting teacher judgments and interpretations of assessment data are fundamental. When discussing whether assessment by teachers could be a dependable option for local and system level assessment, Harlen (2004) recommended five actions to support *dependability* of teacher judgments:

1. careful specification of the assessment tasks;
2. specification of the criteria and standards;
3. teacher training;
4. opportunities for teachers to share interpretations of criteria and standards through moderation; and
5. the development of an assessment community within a school.
In their cross-sectoral large-scale evaluation study of teacher capacity building in assessment, Wyatt-Smith and Bridges (2008) argued that if school-based assessment is to enhance learning improvement within the demands of both system level accountability and local level assessment requirements, careful alignment of curriculum, inclusive pedagogy and assessment must take place. The study investigated the impact of such alignment on the engagement of educationally disadvantaged students in the middle phase of schooling where it was found that … it was important to develop the assessment task and associated standards and criteria at the beginning of the unit: in-depth consideration of assessment before the unit began had a significant impact on the quality of the pedagogy, and thus on student outcomes. (Wyatt-Smith & Bridges, 2008, p. 46)

The study highlighted a concept of front-ending (Wyatt-Smith & Bridges, p.55) the assessment and task design processes as the anchor for curriculum planning and teaching. The authors emphasised that assessment should not be viewed as ‘an endpoint or terminal activity, something tacked on the end of the unit, or done after teaching and learning’ are concluded (p. 46). Instead, fundamental and productive changes in teaching practice can result from critical reflection on the assessment evidence to be collected before teaching begins as this relates to prior learning, as well as during teaching. In the reported study, assessment tasks and associated standards and criteria were developed within a unit of work, each informing the other and proving to establish a clear link between assessment, curriculum and pedagogy. In developing the criteria and standards, teachers drew on their understandings of the local context, the curriculum and literacy demands of the particular year level that they were teaching, their knowledge of the students' levels of achievement and their pedagogical content knowledge of particular subject disciplines. This contributed to powerful teacher development as planning for assessment tasks and defined teacher-generated criteria and standards, in concert with planning units of work, facilitated conversations about quality and learning. In effect, social moderation, discussed below, started at the stage of task design, with teachers interrogating the quality and demands of the assessment they were developing relative to the standards they planned to use for judging quality. Through such a focus on assessment expectations and quality task design prior to commencing the unit of work, the teachers reported they developed a shared language for talking about quality in the classroom and gained confidence in the feedback they gave students.
Before considering Harlen’s third point (teacher training), her fourth point regarding opportunities for teachers to share interpretations of criteria and standards through social moderation is considered. Key to using teachers’ judgments of assessment information for the purposes of local level assessment and system level accountability is ‘a way to integrate teachers’ judgments of students’ responses to the various assessment modes with those of other teachers’ (Wilson, 2004, p. 11). This process is called moderation and is described as a ‘form of quality assurance for delivering comparability in evidence-based judgments of student achievement’ (Maxwell, 2007, p. 2). Maxwell highlighted two functions of moderation namely quality assurance and comparability:

- *Quality assurance* refers to methods for establishing confidence in the quality of procedures and outcomes.
- *Comparability* requires assessment against common characteristics or criteria, such as provided by a subject syllabus or other frame of reference and requires consistency in the application of common standards so that all achievements given the same grade or level of achievement have reached the same standard. (Maxwell, 2007, p. 2)

Several writers (Harlen, 2005; Sadler, 1989; Wyatt-Smith, Klenowski, & Gunn, 2010) have emphasised how common standards provide external reference points for informing judgment and are pivotal for achieving comparability. These writers have argued that it is important for teacher assessors to develop common understandings of mandated standards and reach ‘similar recognition of performances that demonstrate those standards’ (Maxwell, 2001, p. 6). However clear communication about the nature of standards and the levels they seek to specify is not necessarily achieved through the provision of stated standards alone. Sadler (1989) argued, for example, that exemplars or samples of student work provide concrete referents for illustrating standards that otherwise remain abstract mental constructs. He made the point that the stated standards and exemplars work together to show different ways of satisfying the requirements of say, an A or C standard. A related point is that standards written as verbal descriptors call for qualitative judgments. As such the standards necessarily remain open to interpretation and common understandings of the terms used to capture notions of quality in the standards need to develop over time.

While standards and exemplars can serve to make clear desired characteristics of quality, they do not necessarily account fully for the factors that shape teacher
judgment. In a large-scale Australian study of teacher judgment in middle schooling, Cooksey, Freebody and Wyatt-Smith (2007) reported high levels of variability in teachers’ notions of quality and also unearthed the range of factors that shape how judgments are reached. Similarly, (Wyatt-Smith et al., 2010; Klenowski & Adie, 2009) found that while teachers take account of centrally developed and mandated standards, their judgment acts, as displayed in moderation sessions, go well beyond the use of defined standards. Specifically, while they use standards and related textual resources (e.g., Sample responses & Guide to making judgments), they also actively refer to other tacit knowledges (e.g., teachers’ personal knowledge of students; knowledge of curriculum and teaching contexts where they have taught the curriculum; tacit or in-the-head standards not elsewhere specified) for judgment purposes. Parts of this second category of resources were often used in combination, and sometimes pulled against the stated standards. At times, the other knowledges were used as a reason for discounting, or even subverting the stated standards. Given this, it is crucial that practical, unambiguous guidelines and professional development opportunities (such as participation in moderation practice) be provided to teachers about desired judgment practice and the legitimacy (or otherwise) of the various resources available for teachers to draw upon.

This returns us to Harlen’s remaining points namely teacher training and the development of an assessment community within a school. This is explained by Gunn (2007):

Perhaps the clamorous challenge [is] ongoing support for teachers in the form of appropriate professional development and preservice education. Any change to assessment processes (or in fact any educational reform) hinges on support from teachers, and support for teachers, to ensure an ability to adapt to changes at the classroom level. Such changes ideally should follow deep professional knowledge and understanding of the system to be implemented. The challenge is to ensure that teachers have the requisite time and space to reflect on their own beliefs, enter into professional conversations collaboratively across, as well as within, schools and gain deep understanding of the theoretical underpinnings of the various methods of assessment. This requires theoretically-based yet practically-situated learning rather than decontextualised one-shot professional development. (p. 59)

To conclude, quality assurance refers to methods for establishing confidence in the quality of procedures as well as the quality of outcomes. Accordingly the above
discussion outlines a number of quality assurance processes that need to be in place to produce high quality evidence (i.e. valid, reliable and rigorous) of student learning in school-based assessment.

Now turning to an evaluative case study the context in which middle school teachers in Queensland, Australia generated locally relevant assessment tasks for classroom use. Within this context of system level quality assurance for comparability purposes the conditions for promoting confidence in assessment are exemplified.

Case study of the evaluation of the 2008 trial of the School-devised Assessment Model

Queensland assessment context
The Queensland Curriculum, Assessment and Reporting (QCAR) Framework was developed by the Queensland Studies Authority (QSA) with the aim to align curriculum, assessment and reporting for students in Years 1-9 (Queensland Studies Authority, 2008). The focus of the framework is on improvement, particularly in terms of providing clarity related to the key learning area syllabus documents and the consistency of what is taught across the state. The aims of the framework also include providing teachers with support in the form of defined standards and guidelines to help build their capacity i) to make judgments about student work, ii) use information to provide formative feedback to students, and iii) maintain comparability of reported student achievement to parents and carers.

The QCAR Framework consists of five core components: Essential Learnings, Standards, an online Assessment Bank, Queensland Comparable Assessment Tasks (QCATs) and Guidelines for Reporting (www.qsa.qld.edu.au).

QCATs are authentic performance-based assessment tasks. They are administered each year to students in Years 4, 6 and 9. Their purpose is to provide information about student learning in targeted Essential Learnings in

---

1 The research reported in this case study was funded by the Department of Education and Training as part of the Evaluation of School-devised assessment Queensland Comparable Assessment Tasks (QCATs). While all care has been taken in preparing this Publication, the State of Queensland (acting through the Department of Education and Training) does not warrant that the Publication is complete, accurate or current. The Department of Education and Training expressly disclaims any liability for any damage resulting from the use of the material contained in this Publication and will not be responsible for any loss, howsoever arising from use of, or reliance on this material.
English, Mathematics and Science and to help promote consistency of teacher judgments across the state. (QSA, 2010a)

Two models of the QCATs were devised. The first focused on centrally (system)-devised assessment tasks, and the second, on school (teacher)-devised assessment tasks. Information about the centrally-devised assessment tasks is readily available from the QSA website (www.qsa.qld.edu.au). See Figure 1 for the graphic relational overview of the five core elements of the QCAR Framework and the feedback loops of the two models of QCATs.

![QCAR Framework](image)

Figure 1: The relationship among the five core components of the QCAR Framework

In 2010, the identified elements of the framework and the QCATs continue to exist. Some modification to the school-devised assessment model resulted from the external evaluation conducted in 2008 of the school-devised assessment model that aimed to:

- determine whether the school-devised assessment model provided sufficient rigour, validity and reliability to meet the objectives of the Queensland Curriculum Assessment and Reporting (QCAR) policy, and
- provide recommendations for potential improvements to increase comparability and consistency of teacher judgment and the sustainability of the model.
The insights gained, which are central to the reconceptualised model presented in this paper, focus on how confidence in teacher judgment can be achieved in a system approach to the development of comparable school-devised assessments.

**Evaluation methodology**

Participants included teachers, principals and sector representatives including those from the Queensland Studies Authority, Department of Education and Training and the three educational sectors of Education Queensland, Brisbane Catholic Education Commission and Independent Schools Queensland. Of the 17 schools involved in the trial of the school-devised assessment model, six were from the state sector; five were Catholic schools, and a further six were from independent schools. Participation in the evaluation was on a voluntary basis so that each stage of data collection involved all or a subset of teachers from these schools.

Accounts of how the model was implemented were collected from key personnel in each sector who were responsible for providing guidance and support to schools at stages throughout the implementation process. These personnel were key in the evaluation design as they were influential regarding the involvement of the classroom teachers throughout the implementation process.

Data were collected through questionnaires, forums, interviews, documents and observations. The number of specific data collection instruments, the participants and the focus of each method informing the evaluation of the model are represented in Table 1.

**Table 1: Evaluation Methodology**

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Participants</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questionnaires</td>
<td>Teacher Questionnaire 1 Eight questions, varied response options including Likert-scale, open and closed. 18</td>
<td>• validity and reliability  • guiding materials  • development of Assessment Proposal process</td>
</tr>
<tr>
<td></td>
<td>Teacher Questionnaire 2 Eighteen questions, 16</td>
<td>• preparation of the assessment package • participation in the</td>
</tr>
<tr>
<td>Method</td>
<td>Description</td>
<td>Questions</td>
</tr>
<tr>
<td>----------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Panel Questionnaire</td>
<td>Twenty-one questions, varied response options including Likert-scale, open and closed.</td>
<td>19</td>
</tr>
<tr>
<td>Interviews</td>
<td>Interview schedule Ten questions</td>
<td>4 (one from each sector and one from QSA)</td>
</tr>
<tr>
<td>Document collection</td>
<td>Assessment Proposals</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Assessment Packages</td>
<td>19</td>
</tr>
</tbody>
</table>

Each method of data collection was planned for particular scheduled implementation stages\(^2\) and several independent ones. This approach was supported by the Queensland Studies Authority through the dedicated allocation of time for data collection within the running schedule for each of the stages. Additional data collection opportunities were identified during the evaluation.

The design of the evaluation was purposefully constructed involving a staged approach to data collection, which ensured a balance in data collection needs and

\(^2\) Implementation schedule stages refer to the key Quality Assurance Checkpoints and other activities planned in support of the school-devised assessment model trial.
least disruption to participants in the given timeframe from March through to November 2008. Each stage produced a distinctive data set (Table 1) that was analysed individually and then considered collectively with a synthesis of all data sets informing the findings and conclusions.

**The trialled School-Devised Assessment Model**

The model, trialled in 2008 in Year 4 Science and Year 9 Mathematics, comprised two main elements: Element A (*Design Brief*) and Element B (*Four Quality Assurance Checkpoints*).

**Element A – Design Brief**

The *Design Brief* presented a guide for teachers on the Quality Assurance Checkpoints of the model, as well as important information to support them in developing assessments. The stated purpose of the *Design Brief* was to articulate ‘a methodology for collecting evidence of student learning through school-devised assessments’ (QSA, 2008a, p. 3). Specifically, it was developed ‘to provide a set of parameters to maximise validity and reliability in the creation and implementation of school-devised assessments’ (p. 3).

**Element B – Quality Assurance Checkpoints**

The *Design Brief* contained information on ‘quality assurance procedures [aimed at] assuring appropriate and comparable assessment standards have been met’ (QSA, 2008a, p. 4). The quality assurance procedures were to be realised through four Quality Assurance Checkpoints designed to assure that schools develop assessments of quality and apply the standards consistently when making judgments, as outlined next.

Quality assurance procedures ensure that schools have fulfilled accountability requirements by producing assessments which have:

- intellectual rigour, validity, and fairness
- the potential to produce a high quality of evidence of student achievement
- applied the design elements for quality assessment.

(QSA, 2008a, p. 9; QSA, 2008b, p. 8)
While the *Design Brief* used the terms rigour, validity and reliability, also used were the design elements for quality assessment namely credibility, intellectual rigour, authenticity and user friendliness (referred to during the project as CIAU).

The four Quality Assurance Checkpoints as described in the respective *Design Briefs* were as follows:

<table>
<thead>
<tr>
<th>#</th>
<th>Process</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Assessment proposal</td>
<td>An assessment proposal is approved by the principal and then submitted to the Advisory Committee for recording.</td>
</tr>
<tr>
<td>2</td>
<td>School-devised assessment</td>
<td>School-devised assessment packages are approved by the school’s principal and then submitted to the cross-sector panel.</td>
</tr>
<tr>
<td>3</td>
<td>Cross-sector panel feedback</td>
<td>Cross-sector panels provide feedback and advice to trial schools on their school devised assessment package.</td>
</tr>
<tr>
<td>4</td>
<td>Cross-sector moderation</td>
<td>Schools in the trial will meet to discuss their A to E representative samples of judgments made by teachers. The discussions will work towards achieving agreement about the application of standards.</td>
</tr>
</tbody>
</table>

(QSA, 2008a, p. 10; QSA, 2008b, p. 9)

Figure 2 represents the elements of the school-devised assessment model used in the trial, as described in this section. Also represented in the figure are aspects of the model that corresponded with planned training and formal feedback to teachers.
Figure 2: Elements of the 2008 trialled model for school-devised assessments

Essentially, the *Design Brief* provided a guide to teachers of the requirements for fulfilling each of the four Quality Assurance Checkpoints of the model. During the implementation, training\(^3\) was planned for delivery at three specific points in the model. These were: the QSA introduction to the trial, cross-sector panel, and cross-sector moderation. The aim of the training provided at the two Quality Assurance Checkpoints was to facilitate teacher engagement in those quality assurance activities.

Also shown in this representation are two feedback loops, the first of which occurred during the panelling of tasks. Panelling involved teachers and sector representatives reviewing the school-devised assessment materials for endorsement. Panel feedback was intended to inform revision of the task for final submission before implementation in the classroom. The second feedback opportunity was provided at the finalisation of the cross-sector moderation. At this point the focus of the feedback was to review the graded student samples and to evaluate how the standards had been used in the grading process and thereby analysing teacher use of the standards in school-based judgments.

\(^3\) The term training is used here deliberately to underscore that that model called for teachers and sector personnel to undertake assessment planning activities that called for new ways of documenting assessment requirements and using standards in moderation practices.
Findings from the Evaluation Case Study

The analyses of the composite data set showed both strengths and limitations in the trialled school-devised assessment model. The summaries of the analyses are presented in two parts. First, the findings from the evaluation of the elements of the trialled model, that is, the Quality Assurance Checkpoints (QACs) are discussed, followed by the findings relating to the Design Brief.

Assessment Proposal – Checkpoint 1

Overall, the analysis of the Design Brief provided some, though limited, information about the requirements to be met in the assessment proposal. The Design Brief identified two items for inclusion, with the remaining requirements presented as, eight non-negotiables. A lack of clarity about the requirements was reflected in the variability in the depth and breadth of coverage of the outlines of the assessment strategies and rationale statements in the assessment proposals submitted. Results from the audit showed, for example, a limited number of teachers accurately met the intended requirements of the proposal in relation to the coverage of the targeted Essential Learnings (ELs) (15.79%); or specified the assessment strategies used to gather evidence on them (36.84%). The audit also showed high variability in relation to how the schools addressed the second list of eight non-negotiables across the Proposals. Observations and document collection showed that these requirements were not clarified during the initial training day, nor covered in sufficient detail to ensure consistency or comparability in a subsequent meeting.

The requirement for teachers to indicate the varying emphasis placed on the State’s targeted ELs, outlined in the Design Brief, were not elaborated in relation to the scope and coverage. In practice there was high variability shown in the audit of the schools’ assessment proposals. Only three Proposals (15.79%) fully engaged with the targeted ELs and clearly showed the emphasis placed on the ELs in the assessment. Two (10.53%) Proposals did not engage at all with the ELs, with the mid-range being that nine (47.37%) listed all of the targeted ELs but did not engage further with them. This lends support to the conclusion that, while the Design Brief provided some information, it did not fully support teachers in their first attempt to engage with developing comparable school-devised assessment tasks.

4 The findings presents a summary of the evaluation findings across the QACs. Readers interested in more information on the analyses should contact the authors.
The design of valid assessment tasks presented greater conceptual difficulty for teachers than did the practices and processes for achieving reliability. Teachers were uncertain about the relationship of validity to curriculum, assessment task design and other assessment collection strategies. They did not associate validity (as an assessment priority) with how they connected curriculum requirements to classroom assessment tasks. In regard to reliability, teachers emphasised the role of the A to E standards framework, the Guide to making judgment (or equivalent marking schemes) and moderation.

These issues indicate a need for greater specificity in information provided to teachers about task design, with a view to ensuring clearly stated expectations of the approach to be taken by schools to ensure validity and reliability. Further, while the Design Brief required that teachers develop a rationale addressing the design elements for quality tasks (CIAU), teachers reported that this involved new ways of working, especially in connecting curriculum, learning and teaching, and assessment.

**Assessment Package – Checkpoint 2**

The audit of the schools’ assessment packages showed a high level of adherence to five of the submission requirements, with variability evident across the items in the depth and breadth of coverage of information included within the materials. Table 1 presents five of the requirements for the Package with audit findings about the variability documented.

Table 2: Audit findings for assessment package requirements

<table>
<thead>
<tr>
<th>Assessment Package Requirements</th>
<th>Audit Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>the assessment as presented to students</td>
<td>Provided to a high level in Packages. Most Packages presented a booklet identifying the demands of the task that the students were expected to read and to provide a response. Several submissions indicated the use of ICTs as the means for students to present their assessment response.</td>
</tr>
<tr>
<td>clear description of assessment conditions</td>
<td>Sufficient information on the assessment conditions was included in most of the Packages to enable teachers unfamiliar with the assessment design processes to be able to implement the assessment.</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>a Guide to making judgments directly aligned with the targeted Assessable Elements, using the A to E standards framework, and providing an overall grade</td>
<td>All of the Packages included a Guide to making judgments or equivalent marking schema using the A to E standards framework. Variability was evident in these Guides.</td>
</tr>
<tr>
<td>an indicative &quot;A&quot; response</td>
<td>This was provided in most of the Packages. Some variability was evident in the standard of the A response with no indication of whether the illustrative A sample was at the threshold, mid-band or top level of the standard.</td>
</tr>
</tbody>
</table>

The remaining three submission requirements referred to the coverage of the Essential Learnings and the specificity of the Guides to making judgments (or equivalent – referred to now as Guide/s). There was significant variability in the specificity of the Assessable Elements within the Guide, and in the complexity of the Guide, both of which have implications for ensuring construct validity in task design and reliability of teacher judgments. Panellists’ responses (11; 61.11%) to a questionnaire indicated that they did not believe that the Design Brief contained clearly stated expectations in regard to how teachers were to develop comparable assessment tasks and accompanying Guides to making judgments.

Given the variability in the inclusion of the ELs, in task design and related conditions, and the Guides, a key area for teacher capacity building identified was how to use exemplars. As mentioned previously, teachers had limited experience with the ELs, which had only recently been finalised, and with writing task-specific descriptors. These factors may account for the variability. With appropriate training, guideline documents and time, such variability is likely to reduce significantly. The developing Bank of Assessment Tasks (https://qcar.qsa.qld.edu.au/assessmentbank/) is a
valuable tool for achieving this, especially if the assessment instruments are accompanied by other materials, as presented at the end of this report.

**Endorsement process – Checkpoint 3**

A strength of the endorsement process was the approach taken to identifying both strengths and weaknesses of the schools’ submissions. Each Review form first identified strengths and/or commendations before addressing perceived weaknesses.

Requirements for endorsement, provided to teachers, were usually accompanied by suggestions from the reviewers that included implications for action. Teachers reported that they found this feedback useful. Importantly, the tone of the feedback was collaborative and ‘relational’ in tone. Examples of the language used in the Review form include:

- *It is an interesting and motivating task for children.*
- *Thank you for allowing us to read your task. We loved the way it is both individual and collaborative in design.*
- *Thank you for your time and hard work.*
- *The Panel felt that any teacher could ‘pick up’ and teach the task confidently.*

All panellists (questionnaire) reported being satisfied with the process for providing feedback to schools on the Packages. They too provided many comments about the utility of the feedback provided to schools, some identifying the key role of the suggestions:

- *Assessment is positive & specific - does not require teachers to rework without suggestions made by the Panel.*
- *The feedback provided should be useful for schools as it is specific to the task and given recommendations/ suggestions for improvement.*
- *Recommendations/Actions are practical & should easily be incorporated by the schools.*

Domain specific difficulty was experienced in relation to the coverage of the ELs in the assessment task. For example, many comments were made by teachers about the difficulty in incorporating all of the targeted ELs from Year 9 Mathematics into the task. One teacher provided the following comment in a Forum: “Yr 9 Maths ELs - broad selection difficult to embed and integrate”.

---

17
Overall, this Quality Assurance Checkpoint functioned as a quality assurance process and it was recommended that it retain this status (as distinct from Checkpoint). Further, the timeframe for review of the Packages (that is, those with endorsement pending) was reported as reasonable by teachers and the process used for finalisation was acceptable to teachers and facilitators, with one reported teacher exception.

**Moderation – Checkpoint 4**

It is worth noting at this point that the teacher participants had little, if any, prior experience in moderation, and none had experience in applying defined standards in moderation. The outcomes are therefore encouraging. Specifically, following the moderation meeting all judgments of standards (A-E) were confirmed for 14 (73.68%) schools. Judgments were not confirmed in five of the schools as per the following: one (5.26%) school had one judgment not confirmed; two (10.53%) schools had two judgments not confirmed; and two (10.53%) schools had three judgments not confirmed. Overall, there was a high degree of comparability in judgment decisions between individual review and group review.

During the panel forum, participants were asked to rate their confidence that comparability was achieved using a 1 (low) to 4 (high) scale. Of the 28 responses to the question, 18 (64.28%) rated their confidence as level 4, the highest level, and 9 (32.14%) rated it as level 3. Only one response was received with a rating lower than these (level 2). When asked if the assessment tasks viewed during moderation were of comparable curricular demand, 22 (78.57%) reported that they were, while 6 (21.43%) indicated that this was not the case.

One concern of the process for reaching consensus is that the classroom teacher’s direct input on task design, during the cross-sector moderation meeting, may influence the individual reviewer’s judgment decisions. This was highlighted by a comment on one of the forms: *Classification of task requirements from teachers at the school assisted members to discuss and analyse the judgments that had been made by the school.* It was observed by some panels that the teacher providing the input on task design continued to talk with reviewers throughout the individual review period. This directly impacted on review processes and could have influenced the outcome of moderation in this case. The data made clear however that the presence of the teacher was regarded by other teacher participants to be a valued part of the
process. Further, participant comments also indicated that the feedback sheet provided to schools was another opportunity to make suggestions about assessment task design, implementation processes and the use of the Guide in the local context of individual schools.

The cross-sector moderation functioned as an effective quality assurance process that placed the focus on reliability of teacher judgment and reported results. Further, there was strong support from teacher participants for cross-sector moderation (that is, across State, Catholic and Independent sectors), many reporting that it was the first time they had the opportunity to view student samples outside their own schools, and to discuss the bases of judgments with peers.

**Design Brief**

The design brief provided teachers with the details for developing a school-devised assessment and the timeline for navigating the four Quality Assurance Checkpoints. In particular the brief focused on: the design elements for quality tasks (CIAU), the structure and useability of Guides to making judgments, and moderation practices.

As evidence suggests in Quality Assurance Checkpoints 1 and 2, there was insufficient detail in the Design Brief with regards the depth of coverage of the targeted ELs and the requirements for their inclusion in the assessment proposal.

Data analysis revealed that teachers wanted more information in the Brief to support the school-devised assessment process and to ensure that it more effectively supported their efforts. More contextual and background information about the placement of the model in the State’s curriculum and assessment approaches was requested by teachers together with more information about the significance of the Checkpoints. Data analysis of the questionnaire returns and of the forums indicated a lack of understanding by teachers of the validity and reliability demands of the model. The links between the design elements of quality tasks and rigour, validity and reliability remained unclear throughout the duration of the trial.

The Design Brief is a guide that, of itself, cannot provide sufficient training and development for teachers engaging in a process to assure comparability for comparability is influenced by many factors beyond the Design Brief. These factors include the evaluative and curricula experiences of teachers, particularly in the context of a standards-referenced framework. Clarity of the information contained within the Design Brief and consistency of message across all of the communication
channels and mediums enhances the comparability and rigour of school-devised assessments and has implications for future developments.

**Future directions**

In building dependability of school-based assessment and of teacher judgment, Harlen (2004) recommended specification of the tasks and the learning goals with detailed specification of the criteria linked to those goals. In terms of teacher professional development, emphasis was given to those known shortcomings of teacher assessment, such as teacher bias and whimsy (Sadler, 1987). Other recommendations were the practice of moderation to develop teachers' understanding of learning goals and related criteria and the development of an assessment community in which assessment is discussed constructively and positively. In Table 3 these recommendations are built upon by underscoring the factors that unite in realising quality assessment for system reporting as relating to school-devised assessments. Essentially, these relate to information, training, submission points and quality assurance mechanisms as conceptually shown in Figure 3.

**Table 3: Building dependability in the school-devised assessment model**

<table>
<thead>
<tr>
<th>Harlen, 2004</th>
<th>School-devised assessment model evaluation, 2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harlen puts forward five types of action as solutions to increase dependability of teacher judgment and ultimately produce quality school-based assessment:</td>
<td>The current evaluation suggests a number of processes, checkpoints and resources that need to be in place for quality school-based assessment:</td>
</tr>
<tr>
<td>• Careful specification of the tasks and clarity of learning goals</td>
<td>• Clearly specified learning domain (i.e. in the <em>Design Brief</em>)</td>
</tr>
<tr>
<td>• Detailed specification of the criteria that are linked to learning goals</td>
<td>• Resources (PD and time) for planning for assessment tasks and teacher generated Guides/criteria sheets</td>
</tr>
<tr>
<td>• Teacher professional development/training that addresses known shortcomings of teacher assessment (e.g. bias)</td>
<td>• System and/or local level endorsement of assessment plans. Consideration of the full range of standards evident in discipline document specifications (e.g.,</td>
</tr>
</tbody>
</table>
including the 'halo’ effect, steps that guard against unfairness)

- Moderation as a means of developing teacher understanding of learning goals and related assessment criteria – important to protect teacher time for planning assessment, in-school moderation etc

- Development of an ‘assessment community’ within the school in which assessment is discussed constructively and positively

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ELs) need to be woven into work programs, with the assessment plans taking account of the demands of tasks and the need for evidence of the intended skills and knowledge being assessed. A key part of this process is careful specification of criteria and standards at the task level.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Inter- and intra-school moderation practices</strong> to ensure teacher judgments in different classrooms/settings align with each other for consistency of interpretation and implementation of criteria and standards.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Ongoing <strong>professional development</strong> in task development, moderation practices including the social protocols necessary for effective moderation, and knowledge of the legitimacy or otherwise of the various resources that may be influential in judgment.</td>
<td></td>
</tr>
</tbody>
</table>
The first conceptual layer of the model represents the provision of accurate and timely information that may be accessed through a number of mechanisms including the Design Brief, a dedicated website linked to the Assessment Bank, and key personnel supporting the process (facilitators), sector representatives and school leaders including, the curriculum leaders. The second conceptual layer is training that offers support at two submission points (assessment proposal and assessment package) and the two Quality Assurance Checkpoints (endorsement and moderation) (see Figure 3). More specifically, training on the protocols for moderation and the roles of panel members and panel facilitators are required. This dual focus sharpens the accountability checks. Finally, the Assessment Bank provides a clearinghouse for promulgating quality assessment tasks and stated standards.

The system-level approach to building a sustainable assessment culture recommended offers another option of the school-devised assessment model. That is the incorporation of the opportunity for schools to submit their school-devised assessment tasks for accreditation and registration on the Assessment Bank. Such a quality assurance check would help to address any misalignment of curriculum and assessment, when a school-devised model of assessment is implemented at the
same time as curriculum reform. Such misalignment was an issue for teachers in this evaluation. By retaining a model of school-devised assessment that draws on online resources, the opportunity exists to build the teaching profession’s assessment capabilities in using standards. This again offers an opportunity to develop teachers’ assessment skills and understanding during the implementation of curriculum and assessment reform such as currently being experienced in Australia with the Australian Curriculum and achievement standards.

Teachers’ use of standards at both task and discipline levels ensures relevance for both national and state curriculum and assessment reform priorities. This strategy involves three components that will contribute resources to the Assessment Bank. In this way further support would be offered to teachers’ work in developing school-devised assessment tasks for accreditation and registration within the Bank.

Drawing on the work of Smith (1989), and Wyatt-Smith and Klenowski (2008), the components for a strategy to support teacher use of discipline standards include:

- elaborated guidelines about on-balance judgment processes, focusing attention on how teachers consider the qualities of the work against each of properties of the standards (e.g., A-E) and how they analyse the configuration of the properties to determine those that are dominant in the student work

- exemplar student work (on a task, extended to a portfolio) indicative of the standards, illustrative of a particular achievement level (A-E). While these could be located as within-band level, they could be more usefully chosen to illustrate the absolute minimum requirements for work judged to be a particular level. Such threshold level exemplars would be particularly useful to illustrate the minimum requirements for a particular level such as a C level. The role of these materials is to illustrate different ways of satisfying the stated requirements of the standards

- descriptive reports of student achievement accompanying the exemplars to give insight into the factors that influence the overall judgment and the final achievement awarded. Such reports provide information about the decision-making processes of the teacher in the arrival at an overall judgment, including specifics about how trade-off of perceived strengths and limitations occurs.
Conclusion

There is growing evidence in published research and practice in countries such as Canada, New Zealand, England, Scotland, Wales, Hong Kong and Singapore that schools are adopting distributed leadership, with the role of the principal reframed as leading in learning. Also evident is that governments continue to call for evidence of the quality of schools and the impact of school improvement initiatives. It is in this historic context of a new century with challenging and unpredictable demands that educators face in sharp relief the dual focus on learning, with system and local efforts to improve learning on the one hand and education policy priorities to generate assessment information on the health of school systems, on the other. In this context there is a real danger that large-scale tests may appear the most tantalising option for governments hungry for information about school performance. This is not suggesting that large-scale tests do not have a role to play. However, a call for intelligent accountability that recognises how, under certain conditions of system checks and balances, quality classroom assessment can achieve the purposes of learning improvement and reporting. In so doing, more dependable and sustainable assessment cultures are possible.

Acknowledgements

Our sincere gratitude goes to key staff in the Department of Education and Training, Queensland Studies Authority, Education Queensland, Brisbane Catholic Education Commission, Independent Schools Queensland, and the principals and teachers who participated in the evaluation. Without the participation of these key stakeholders the evaluation would not have been informed by the diverse views held by these groups.

References


---