

WYA07015
AARE 2007 International Education Research Conference Fremantle
Research Impacts: Proving or Improving?

Fremantle
Western Australia

(25-29 November, 2007)

‘Moderation as Judgement Practice: Reconciling System Level Accountability and Local Level Practice’,

**Klenowski, V., Adie, L., Gunn, S., Looney, A., Elwood, J.,
Wyatt-Smith, C. and Colbert, P. (2007)**

Abstract

This paper is based on the very initial stage of a four year research project focused on teachers’ judgements and interpretations of assessment data in the context of standards-driven reform in Queensland. This major research project is investigating the local level practice of social moderation as a means of developing shared understandings of standards and consistency of judgement amongst teachers in face-to-face and ICT contexts for the purposes of demonstrating system level accountability. The authors draw on a social theory of learning (Wenger, 1998) and assessment research to illustrate the importance of acknowledging that teachers are the primary change agents who, through judgement practices, are best placed to identify important steps for students to improve in their learning. The initial stage of this research reported in this paper suggests that the practice at the local level of social moderation has the potential to fulfil 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 so doing attend to system level accountability.

‘Moderation as Judgement Practice: Reconciling System Level Accountability and Local Level Practice’

**Klenowski, V., Adie, L., Gunn, S., Looney, A., Elwood, J.,
Wyatt-Smith, C. and Colbert, P. (2007)**

Introduction

Internationally, recognition of the role of assessment to inform the learning process has received much attention in recent years. Assessment *for* learning, not just *of* learning is being supported by an increasing body of literature providing strategies that teachers and their students can incorporate to support the learning process (Assessment Reform Group, 2002; Broadfoot & Black, 2004; James, 2006). Concurrently there has been an increase internationally in systemic accountability requirements of schools in terms of student results. The convergence of these two movements has resulted in some education systems promoting standards-driven reform involving authentic assessment and a re-examination of the relationship between the teacher and the student in the learning process. In this context standards are intended to be used as the basis for judgements of student achievement; while the results from assessment tasks are meant to both inform the teaching/learning process, and to report and track student progress. In such a system, the role and reliability of teacher judgement takes centre stage.

This paper reports some issues that have emerged in the preliminary stage of this research project (funded by the Australian Research Council [ARC]) being conducted in the Australian State of Queensland. The focus is on the use of state-imposed standards to promote consistency of teacher judgement of student work. The major study aims to examine whether social moderation, as an explicit form of judgement practice, progresses the development of teacher agency. Teachers are best placed to identify important steps for students to improve in their learning and to develop useful insights about how best to change pedagogy to meet students’ particular learning needs. For the first time in Queensland, teachers in middle and primary schools are using defined standards to form judgements of student work and are engaging in constructive and professional debate about judgements of students’ work by other teachers. The development of an ‘assessment culture’ within schools and across the state that is integral to the development of quality learning opportunities for students is anticipated. It is also intended that such engagement will lead to enhanced teacher capacity for assessment.

Rationale

The last decade has witnessed an increased use of international and national tests to monitor standards of achievement in schools and to report on results for public accountability purposes. An unintended consequence of this form of testing has been that teachers collect test data rather than use that data to inform teaching and learning (Black & Wiliam, 2005). International concern about the utility of testing for monitoring standards of achievement in schools has renewed interest in the links between standards, school-based teacher judgements and quality learning (Wyatt-Smith, 2005).

Since the late 1990s in Australia, national standards (in the form of minimum benchmarks) have been used for reporting student literacy and numeracy performance. Students are involved in international testing and comparison measures (for example, PISA and TIMSS) as well as state literacy and numeracy tests in Years 3, 5, 7 and 9 (national tests in 2008). The assessment data is used for accountability purposes at the national level to inform schools and parents about the performance of a cohort or an individual student in comparison to the national or state average and at the school level to identify areas of strength and weakness in a program. These high stakes tests are externally marked with a substantial gap between the time of the test and the distribution of results. Consequently, teachers and schools view them as accountability measures and do not consider them as useful for informing and involving students in their learning process.

The rationale for giving more consideration to teacher judgement practice within a standards-referenced framework is the link between assessment and improved learning for students. The provision of quality standards can inform decisions at the central political level and can assist in meeting the demand for public accountability at the local professional level of the teacher (Wilson, 2004; Harlen, 2005a). Standards can make explicit for teachers what to teach and the level of performance expected for a particular age group. Research indicates that standards are useful for informing teachers' work and in contributing to quality teaching and learning experiences (Wyatt-Smith & Castleton, 2004; Sadler, 2005; Klenowski, 2006, 2007). Yet teacher judgement remains largely uncharted territory and influences on judgement need to be investigated (Wyatt-Smith, Cumming & Elkins, 2005) particularly if the education system is committed to "characterising the teacher as the key person in the assessment process ... [with] teacher judgement at the heart of that process" (Maxwell, 2002:13).

Current research suggests there is significant variability in teacher judgement and teachers' underpinning conceptualisations of assessments (Elwood & Klenowski, 2002; Cumming, Wyatt-Smith, Elkins & Neville, 2006). In Australia there has been a lack of consistency and accountability in assessment and judgement practices across the middle years of schooling. Teachers appear to have varying levels of knowledge about quality assessment and the use of standards to inform judgements (Luke, 2003; DET, 2005). Teachers' assessment capacity needs to be developed, specifically in preschool to Year 10 (students aged 5-15 years) (Education Queensland, 2002:3). This context for change emphasises the importance of validity and reliability issues associated with the incorporation of social moderation as a means of informing and improving teachers' judgement practice.

While research on different models of moderation and their impact on judgement consistency is limited (Ingvarson, 1990; Malone, Long & De Lucchi, 2004), some studies (Hutchinson & Hayward, 2005) have identified moderation as an essential support for teacher judgements. Other studies have shown that the provision of a standards-referenced framework does not of itself regulate judgement, nor does it make explicit the factors that shape judgement (Maxwell, 2002; Harlen, 2005b). From the perspective of a social theory of learning it is important to emphasise that the statements of standards that are being introduced to teachers for the first time will not be understood in the same way

by all teachers; rather it is anticipated that there will be multiple meanings. This study will assist our understanding of how a standards-referenced framework shapes teacher judgements during social moderation and the dynamic processes of judgement and decision-making or the 'negotiated meaning' in group settings.

Queensland is an interesting context for this research as the number of small and isolated schools and the distances separating them, present major challenges for face-to-face moderation. Given these factors, new practices using different modes and media to facilitate moderation processes, including working with portfolio evidence (Klenowski, 2002) are being explored. In this study face-to-face models of moderation together with ICT-mediated moderation environments that afford virtual moderation meeting spaces are being trialled. The focus for the ICT-mediated moderation is on how such spaces enable moderation via online 'communities of practice' (Lave & Wenger, 1991; Wenger, 1998; Slack & Hine, 2000; Stehlik & Garden, 2005; Wise, 2005). The technology is also used to capture screen and talk data with high levels of accuracy. The paucity of data about moderation practices especially in remote and rural areas of Australia and an absence of sustained study on the efficacy of ICT-mediated moderation have provided avenues for research.

Background and Context

Queensland Senior (Years 11 and 12) syllabuses and the assessment procedures used to determine exit levels of achievement for the purposes of issuing a *Senior Certificate* are standards-based. For the past twenty years, Queensland senior school teachers have been working in a criteria-based standards-referenced assessment system. Within this system, teachers are required to write comprehensive and detailed school work programs which specify various aspects of a course of study that they have tailored in response to mandated syllabus documents. These include: the variety of learning experiences to be undertaken; the available resources necessary for satisfactory completion of the course; the range of assessment items and the conditions under which they will be undertaken and the means by which the teachers will determine grades on course completion. In practice, standards-based assessment requires that teachers prescribe and publish criteria and standards prior to students commencing work on an assessable task. Routinely, explicitly stated standards with accompanying assessment criteria are used by teachers to make judgements about the quality of student achievement as evidenced in folios and recorded in profiles. With this history of school-based curriculum and an approach to externally moderated standards-based assessment for the senior phase of schooling (Years 11 and 12) it is considered that consistency of teacher judgements has been developed for this phase.

However, this history does not extend to the middle years of schooling in Queensland because of the constraints of distance (approximately 1.9 million km²) and the number of schools (approximately 1700). It is neither logistically nor fiscally possible to replicate the senior model that applies to 280 senior schools and for two years of schooling. This has resulted in a lack of clarity in assessment and reporting in the early and middle years. In these years there is a wide range of curriculum planning, assessment and reporting practices and no formally endorsed assessment and reporting framework. To align curriculum, assessment and reporting, the Queensland Government has developed the

Queensland Curriculum, Assessment and Reporting (QCAR) framework. As part of this reform the Queensland Studies Authority (QSA) is developing and trialing standards for each Key Learning Area (KLA) at year level junctures 3, 5, 7 and 9, for state-wide implementation in 2009. Standards (A to E) are being developed to inform teachers' judgement-making about student achievement for local and system reporting. This initiative was taken concurrent with the Australian Government's move in 2006 to implement an A to E reporting system for the purpose of providing standards-related information to parents.

The QCAR framework consists of five components;

- Essential Learnings, which describe the key knowledge, facts, procedures and ways of working for a particular KLA by the end of a year level juncture (Queensland Studies Authority, 2007a);
- Standards for the Essential Learnings, which describe the expected quality of student work, and provide a common language for teachers to use in discussing student work (Queensland Studies Authority, 2007b);
- An on-line assessment bank, which will provide examples of quality assessment tools, and include exemplars of student work, and guidelines on using the standards;
- The Queensland Comparable Assessment Tasks¹ (QCAT/s) for Years 4, 6 and 9; and
- A reporting framework, which encompasses advice to schools on their twice yearly reporting to parents as well as the reports specific to the QCAT tasks.

The QCAT focuses on a targeted set of Essential Learnings which remain the same for two years though the task itself differs each year. Schools are given an eight week period in which to implement the task, to grade it and to moderate results. Teachers are provided with:

- Copies of the student task, including a guide to making judgements which includes the assessable elements² of the task and the task-specific descriptors³ on an A to E scale;
- Teacher guidelines, which provide recommendations on implementing the task, advice on grading the tasks, and models of moderation that may be used to improve consistency of judgement formation;

¹ The term student assessment task also refers to Comparable Assessment Task as used throughout this paper. Initially the task was to be termed a Common Assessment Task however the change occurred after a meeting with the Board that has representation from each of the three sectors (Independent Schools, Catholic Education and State Schools).

² 'Assessable element' was introduced as an alternative for the term criteria with the intent of assisting teachers to focus their judgement-making using the particular element or quality. It was also introduced in an effort to distinguish the Curriculum, Assessment and Reporting framework which has been designed for middle schooling from that used in senior secondary school model and associated with a different model of moderation. The framework for middle schooling is still in a developmental stage and was being piloted in this initial phase of the research.

³ The term 'task-specific descriptor' refers to the standard as used in the context of A-E standards for the assessment of the task. The terminology was changed from standard to 'task-descriptor' again to avoid association with the senior secondary school model of moderation.

- Annotated samples of student responses, which include A to E responses for each question of the student task.

Standards-referenced assessment and moderation processes aim to promote consistency and reliability of judgement formation. For middle school teachers these are new practices. The major research project is examining how these teachers use the mandated standards to judge the quality of student work and the models of moderation to support consistency of teacher judgement to inform standards-based reporting to students, parents, carers and systems. The utility and cost-effectiveness of the models of moderation for their responsiveness to different pedagogical and geographical contexts and student diversity is also being investigated.

The form of moderation being investigated is otherwise referred to as ‘social moderation’ (Linn, 1993) or ‘consensus moderation’ and involves groups of teachers meeting to discuss and negotiate assigned gradings of student work with the aim of reaching consensus and a common understanding of the quality of work. Matters (2006: 2) defines this form of moderation as “a *set* of processes designed to ensure that standards are applied consistently across teacher-assessors and across schools”. Moderation in the QCAR trial is not mandated but recommended by such terms as sharing or facilitating dialogue around the tasks. The models suggested by QSA are termed the ‘calibration model’⁴; the ‘conferencing model’ and ‘the expert model’ (Queensland Studies Authority, 2007c).

- The calibration model involves teachers individually grading samples of student work from different classes/schools and then discussing these gradings with colleagues. The aim is for teachers to reach consensus using the assessed work samples and in so doing developing a common understanding of the standards. Reaching this understanding occurs prior to the teachers grading all the task responses from their own classes.
- The conferencing model⁵ involves teachers grading some or all of their students’ responses individually. They then select assessment samples representative of their application of A to E grades. They meet with other teachers to discuss their judgements by sharing their samples. Teachers reach a consensus on the interpretation and application of the standards.
- The expert model⁶ requires teachers to mark all student work and then to submit selected samples to an ‘expert’ who confirms whether there is consistency in the way the standards are interpreted and applied. (Queensland Studies Authority, 2007c).

It is anticipated that over the duration of this research project moderation will occur intra-school, interschool as geographic clusters or like-school clusters (for example, the Christian schools) or as ICT-mediated moderation with schools that are in diverse locations across Queensland. A potential issue that is likely to emerge, as a tension, is

⁴ The term calibration was used by QSA to describe this model but calibration is not used in association with statistical moderation in this context.

⁵ The issues reported in this paper emerged from an analysis of the conferencing model.

⁶ The expert model was not observed in the initial phase of this research project.

the variety of interpretations of what counts as moderation. At this stage it is too early to identify if there is, or is not, a shared understanding of social moderation – its purpose and practice.

The prospect of developing state-wide consistency of judgement formation based on stated standards, presents the opportunity to explore different means of gathering teachers for social moderation purposes. ICT-mediated moderation has not been previously trialled in Queensland. This mode while presenting opportunities to defy restrictions of time and space between people, also presents many challenges related to the technology and the establishment of an online community. ICT-mediated moderation is being trialled using the WebEx[®] meeting centre. The research team has been involved in establishing technological and procedural guidelines for these online meetings. The use of technology to facilitate online learning and meetings is a fast growing industry, yet still vastly under-explored. At this stage approximately five out of a possible twenty-five schools have expressed an interest in the initial trial of ICT-mediated moderation for the next phase of data collection. We can only surmise that IT mediated moderation is presenting a simultaneous demand on teachers as they familiarise themselves with a new form of assessment task.

Theoretical Underpinnings

The study is based on a social theory of learning (Wenger, 1998) and a socio-cultural perspective of assessment. Learning is viewed as active construction of knowledge in which the social and cultural environment is integral (Elwood, 2006; Klenowski, Askew and Carnell, 2006). Socio-cultural and situated views of mind and views of learning in the context of assessment frameworks are significant in this active view of knowledge construction. If we are to understand learning through assessment then we need to focus on the interactions between “... the learner, the teacher and the assessment task in the social, historical and cultural context in which it is carried out” (Elwood, 2006: 231). The acknowledgment of the social, historical and cultural contexts in which learning and assessment takes place, requires consideration of the power relations and the values inherent in the interactions between the various contexts ranging from the political, the professional, the curriculum and at the levels of the school, the classroom, the individual teacher and the student. This reform is being implemented in a political climate that moves beyond the improvement of student learning. This socio-cultural and socio-political environment is important to understand as we determine how teachers make meaning of the standards. The social and cultural knowledge and experiences that impact on that understanding and the manifestation of these in the judgement and moderation processes are important considerations in this research.

Socio-cultural conceptions of learning impact on assessment practice so that a sense of community becomes fundamental to the learning process. Lave and Wenger (1991) propose that learners gain experience through a ‘community of practice’ which involves a form of apprenticeship through a process of ‘legitimate peripheral participation’. In this study although there are experienced teachers and those who are new to the profession, the majority are for the first time using the standards to assess the QCAT. This is a new experience and they have had limited opportunity to experience the models of moderation presented by QSA. In this respect they are in many regards all apprentices. This research

is examining the beginnings of the development of a particular assessment community in the context of social moderation.

According to Wenger's social theory of learning (1998:5) there are four components: "community (learning as belonging); identity (learning as becoming); meaning (learning as experience) and practice (learning as doing)". The 'community' in this study is an 'assessment community'; the 'community' of Queensland teachers who through their experience of moderation negotiate an understanding of standards that underpin their judgement formation of student work. This involves the teachers developing an 'assessment identity' as a member of this community through moderation 'practice' where through participation a recognizable level of competence is established. Wenger's four components involved in learning as a social activity underpin the analysis to illustrate the different dimensions in the development of an 'assessment community'.

ICT-mediated moderation provides another challenge for establishing 'identity' and 'community' in an online environment that involves oral but not visual communication between participants, and introduces new community 'members' - the computer, the telephone and associated technology. The different ways of interacting and forming a community that are involved in a network environment have been identified in recent research (Hara & Kling, 2000; Arnold & Smith, 2003; Goodyear, Banks, Hodgson & McConnell, 2004). This project is considering how the information and computer technologies contribute to the shaping of assessment communities and the negotiation of meaning framed within the components of Wenger's theory of social learning. Our aim is to view the inclusion of the information and communication technologies as another 'participant' in this interaction, reflecting the diverse and complex relationships that impact on the development of an 'assessment community' and the shared understandings of the standards.

Research Questions and Design

The research project is designed with two key understandings. The first relates to student diversity. It is anticipated that by monitoring teachers' moderation discourses, including those in rural and remote areas, insights that could leverage improved learning outcomes for a wider group of students will be provided. The second understanding relates to issues of judgement and the view that moderation is a contextualised social practice that can be enacted in a variety of modes and media.

There are three related inquiries connected to each of the key research questions. Inquiry one is focused on the judgements of different assessment evidence and relates to the questions:

- How do stated standards work to inform and regulate judgement in different curriculum domains?
- What are the configural properties of teacher judgements about the standard of the work assessed and how are they explicitly shared between teachers?

Inquiry two studies the development, trial and evaluation of moderation modes and practices with the focus on the research question:

- Does the social practice of moderation involving the application of explicitly-defined standards result in changed judgements about students' work?

Inquiry three examines the portability of the standards across the sites and learning domains by considering the research question:

- Does moderation using standards result in consistency of teacher judgement?

A mixed methods approach incorporating quantitative and qualitative methodologies is being used. The major study will use surveys to identify trends in terms of the models of moderation used and the nature of the samples of student work assessed against the standards. Interviews will be carried out pre and post moderation meetings to seek teachers' views on their understanding of moderation and how they have made use of the standards in their judgement of student work. Teachers' observations about how moderation assisted or informed their judgement of student work, and their understanding of standards, will be sought post moderation. A further data set is the conversation data that will be captured during the moderation meetings and will then be transcribed and analysed. The focus is on the factors that influence teacher judgements.

At the very outset of this research an initial survey of 186 respondents from State (75%), Independent Schools of Queensland (12%) and Queensland Catholic Education Commission (10%) was conducted and the data analysed. The majority of respondents were from primary schools (62.4%) with the remainder from secondary schools (16.1%) and a combination of schools providing P-12 and P-10 (16.1%). The respondents represent evenly the geographical locations: rural (37.7%), regional (33.0%) and metropolitan (26.5%). This survey aimed to gauge teachers' initial perceptions and understandings about the purposes, nature, function of moderation, their plans for and experience of moderation and their perceived support needs. The respondents are not the same teachers who have been interviewed or who participated in the moderation conversations during this initial data collection phase study so this data has *not* been used in the analysis of the emergent issues that are reported here.

This paper reports on the issues that have been identified from an analysis of the interviews with fifteen focus teachers conducted pre and post moderation at a one-day school cluster moderation meeting. There were four one-day cluster meetings organised and managed by QSA. The focus teachers (15) were either nominated by their principal or Head of Department or volunteered to be involved.

Table 1 Composition of the group of focus teachers.

Gender	School Level/KLA	Sector
12 Female	Primary 8 Year 4 (English) 2 Year 6 (Maths)	11 State
3 Male (Secondary)	Secondary 4 Year 9 (Science) 1 Year 9 (Maths)	4 Independent

The conversation data was also collected at the one-day cluster moderation meetings from the focus teachers of Years 4, 6 and 9 English, Science or Mathematics KLAs. Three of the four clusters were from the state education system while the other cluster comprised the Christian schools. The clusters varied from 4 to 12 schools per cluster with a focus on a particular KLA, the researchers observed and audio-recorded moderation sessions following 3-4 focus teachers per moderation cluster.

Prior to the day of moderation, information was sent to the school organiser to distribute to the teachers. This information included ethical consent forms, details about the research project (Appendix One), the program for the day, the models of moderation and information statements on the Essential Learnings and Standards which were also available on the QSA website.

This paper draws on data from QSA documents and some of the data sets (interview and the conversation transcripts that pertain to the Year 4 English Task). A socio-cultural theoretical perspective was used in the analysis and the constant comparative method was employed to identify the issues. First, the research team divided the task of analysis amongst the members so that each data set was analysed as a separate set by one or more of the researchers. After the analysis of the interview data had been completed, and emergent issues were identified and discussed, a partial cross comparative analysis took place.

Emergent Issues

At this preliminary stage it is important to emphasise that the issues that have emerged can only be described as tentative. The data that has been analysed was collected from only fifteen teacher interviewees and from the professional conversations (only Year 4 English QCAT) conducted in the trial of two models of moderation organised and managed by QSA. Essentially these meetings were an opportunity for participating schools to engage in professional development designed to raise awareness of the processes and procedures for moderation practice. At this stage then only the conversation data related to the Year 4 English QCAT moderation meetings was analysed to identify if there was evidence that related to the emergent issues identified from the analysis of the interviews.

Differing use of terminology

What appears to be emerging is differing use of the terms ‘criteria’ and ‘standards’. From the analysis it appears that these terms are being used interchangeably. Sadler (2007) has recently identified that the particular meaning of such terms used in discourse on assessment and grading need to be clarified particularly if the meaning is not immediately apparent from a given context. This is an important issue as teachers require a common language and way of talking about moderation practice for if they are to practice moderation then there needs to be “shared historical and social resources, frameworks, and perspectives that can sustain mutual engagement in action” (Wenger, 1998: 5). If the terms are recognisable and meaningful to participants this will facilitate

and promote professional conversation for meaningful judgement practice and for the teachers' learning.

In the Year Four English QCAT (see Appendix Two for the guide to making judgements) the term assessable element/s (hereafter, criteria) is used to refer to the properties or characteristics of the report genre and report writing which is the focus of this task. In an effort to focus judgement-making the term 'assessable element/s' was chosen rather than the term criterion/criteria. However, in making this change in terminology with the intention of assisting teachers to think about the 'assessable element' to focus their judgement on the 'element' there is a possibility that distinctions in meaning of some terms such as criteria and standards may be lost. Sadler (ibid) has identified that the term criteria has two meanings: "properties" or "a fixed level above which something qualifies for a particular designation, as one 'meets the criteria' for an award or grade" (p. 388). It is important in the practice of moderation to stipulate the intended meaning of terms if this is not apparent from the context.

For each QCAT, task-specific descriptors were provided at the particular levels of A to E as a developmental scale intended for use as reference points. In addition, teachers were given a booklet of five annotated student work samples illustrative of how each product of each QCAT was representative of an A – E grade. From the preliminary analysis it is possible to identify different ways in which teachers were using the criteria and standards to form a judgement about each product.

The Year 4 English Task, which presented students with two different types of text (a cartoon and a newspaper article) and comprised four products, is used in this paper to illustrate some of the emergent issues. The student task booklet described the products thus:

Product 1: What is it all about? - Use information from both texts to fill in a concept map.

Product 2: How do they compare? - Think about what makes texts the same or different. Answer the questions to show your understanding, and report which answers are the same or different.

Product 3: Report for the school newsletter – Write a report for parents and a recent class event.

Product 4: Reflection on your work – Think about how you made your report interesting and believable.

A variety of ways for making a judgement

The teacher guidelines advised teachers to make a judgement about the quality of each assessable element (criterion) and to match the evidence for each to a task-specific descriptor (standard). It was emphasised that they should consider all descriptors before deciding upon a match. In making an overall judgement teachers were again advised that when "there is uneven performance across the assessable elements, trade off better performance in some elements against lesser performance in others to arrive at an 'on-

balance' overall grade judged in the context of the wider intentions of the task" (QSA, 2007d: 8-9).

Teachers used the criteria and standards in a variety of ways to make a judgement about the overall grade awarded. In reaching a decision teachers were advised to consider all the products completed in the task presented for moderation. Some teachers interpreted the standards as the *minimum* required to achieve the particular grade and in making a judgement they appeared to rely on a rather specific interpretation of each word in the standard descriptor. Others interpreted the standards as was intended, that is, as an indication of *typical performance* required to achieve the particular grade. These teachers made a judgement according to 'best fit' or 'best match'. As the interview data also suggested teachers were unsure whether to use the standards specifically, 'to the letter', or as a rough guide to inform judgement. It was apparent from the interviews that length of teaching experience seemed to create differences in how teachers applied the criteria. Those teachers who were more experienced appeared more confident and were more likely to use the standards as a guide while 'the novices', those with less experience, tended to use the standards to the letter.

Varying approaches to the process of moderation

The following transcribed conversation is of teachers from one of the four schools, participating in the conferencing model of moderation for the first time. They are considering a completed Year 4 English QCAT that had been assessed by another school as grade A. The conversation illustrates the approaches adopted and provides evidence that teachers are employing different responses and processes to those suggested in the QCAR documents. The guidelines suggest that teachers consider each product prior to making an overall judgement. Here the teachers consider sample number 11 and decide on the overall grade first.

F1⁷: Can you tell us what the um, school gave the student sample (Number 11) as an overall grade?

M2: Um , A.

F2: It has a question mark.

F1: A, with a question mark. So basically what we are doing is not going to go and grade each of the products, just the overall grade. So, what was your overall grade?

F2: I would tend to agree. I would have given this a B+ but like I said just before, because I wasn't sure I would go up and give them the A instead.

⁷ (F1 is Female 1 and M2 is Male 2 etc)

F3: I did a B. I actually thought there was (sic) a few incorrect things that they ... [indecipherable] so the information was there but it was incorrect in detail, so looking at that in the specific descriptors, I thought that fell into a B.

F1: M1?

M1: Yeah, I gave it a B as well, but I like F2's view of just ... instead of being stuck in an area, if you think it can be better take it up to the A. So you could say A, she's got the A question mark, hasn't she?

F1: ...I gave it a B, only because I agreed that a couple of the internal tasks were incorrectly identified, going by the criteria sheet I could see how she marked but when I cross-checked it with the samples that we had, the last one in particular ... the last task, I actually made it a D instead of a C which brought the mark down to a B, but it was very close ... it was B+. Did anyone else want to comment on sample 11?

The teachers then moved on to consider sample 12 (see below) at this point the facilitator (F1) realised that she had made an error regarding the procedure adopted in the moderation process. That is, that they were expected to consider their individual marks for each of the products then consider those given by the school. The group decided to return to sample 11 at the end.

Comparing across samples

In the following transcribed conversation some teachers showed a tendency to make comparisons across samples despite the advice given that "making judgements is not about determining whether the work of Student X is better than the work of Student Z." (QSA, 2007d: 8).

F1: OK, sample 12. Have you got 12?

..... F2: I think they gave her an overall B, didn't they?

F3: But comparing that with 11⁸ ... I found that much better than 11, I mean she's giving 11 an A so maybe I'll bring that [sample 12] to an A.

The above excerpt from the conversation data, and that which follows, raises the possibility of teachers comparing responses of particular students rather than focusing on the properties or criteria addressed in each individual completed student task. The facilitator continues with the conversation focused on student sample number 12.

F1: Alrighty, with a B, C, and three A's the school rated them [sample 12] a B. Did anybody change their overall mark?

F3: No, until we started talking about it, then.

⁸ Here the reference is to work sample 11 as discussed in the previous section.

F3: ... until we compared. See, I found even when we were doing this that even though you know, with one of these levels here ... even though they met the descriptors and requirements that there are huge levels within each grade level. Did you find that? Sample 11 could be an A and this sample [12] we've decided could possibly be an A. There are huge differences within an A.

M1: Yeah, like an A+

It could be interpreted that what is being compared here are the qualities that constitute an 'A' grade. However, in analysing the conversation related to sample 12 for this group of teachers there were no *specific properties* that were identified and discussed. There was also no discussion about *the extent* to which sample 12 matched the criteria in terms of specific qualities.

Weighting criteria

The notion of giving some criteria a greater weighting than others emerged as an issue in the analysis of the interview data. Such differential interpretation and application is not in the guide to making judgements issued by QSA. Teachers talked of applying different weightings to certain criteria, given their subject knowledge and having ideas about 'higher level skills' being required to meet certain criteria. It appears that this has emerged as an issue because of personal values attached to certain criteria. It tended to raise itself more clearly in the moderation of English tasks. It is one to explore further if subject pedagogical experience means that teachers are applying weightings, explicitly or not, to students' work, but not doing it consistently. To illustrate this issue of differential weighting of individual criteria when making an overall judgement, from the interview data, it appears that teachers often weighted criteria in their own minds to arrive at an overall judgement of the standard.

...the child might have gotten a C and then an E and we were trying to decide an overall mark, are some more heavily weighted than others? ...That's the hard thing.

... a reflection's not as important as a narrative as such...

It was also apparent that some teachers are valuing certain qualities to identify good students or expecting the student to do well on certain criteria for the award of an A grade. This is an issue when teachers use some items on a task to identify 'good' students which constitutes more of the teacher's expectation rather than what the criteria and standards across the grades are 'asking' for. It is also an issue when teachers are spotting for these items and students' responses on them. In other words, teachers are looking for particular responses that reflect their own ideas as to what constitutes a good or an A grade response which may or may not correspond to the criteria for that product.

In the excerpt from the conversation data that follows the property of ‘reflection on one’s work’ is something that is being debated in terms of whether a different weighing should apply to this quality.

F3: And do some things weigh more? Because like ... I didn’t find a reflection [Product 4] as important as writing ...

M1: Yes

F3: A report ...

M1: But then reflection is higher up ...too, so you’d think that’s a sort of higher order skill, but you hardly do it.

F3: But because it’s the end of the task too [Product 4] ... it’s just to get it over and done with.

M1: Plus they only had 10 minutes.

From the interview data it became apparent that some teachers disregard the standards and base their judgements on their insider knowledge of the student and the effort that the student has made. However, in so doing, these teachers are making judgements that threaten the validity of the judgement or in Messick’s (1989) terms this is an example of ‘construct irrelevant variance’. That is, the teacher’s knowledge about the student is impacting on the teacher’s judgement rather than limiting the focus of the judgement to the properties as reflected in the completed work. For example, this interviewee expressed her judgement practice as being reliant on insider knowledge of the student:

... for me I still have to go on what I know of the children..

And another interviewee suggested:

...we use our prior knowledge of what we would expect for that sort of answer and that’s how we determine whether the student was right or wrong.

There was further evidence in the conversation data of the belief that knowledge of the student could impact on the grade awarded, for example:

... M1: Or did the teacher mark it to an A because they knew what the student was like?

... F2: I wouldn’t put that as an A.

M1: What if it was your top student? No I just wonder if people have done that.

M2: My top student got a B in this thing, he misread the task.

M1: All my students got A's.

F2: In proper teaching you don't take that in isolation. If it was my top student I'd give it a B but I would be rating it against other marks for their report card and they might have 3 other A's.

Varying responses to annotated work samples

It is evident from the interview data that for some teachers the annotated examples provided explanations of the criteria according to the A-E scale. They used the annotated graded samples of student work as a resource in forming their judgements while others, as evident in this excerpt from the conversation data found the annotated work samples less helpful.

F2: ... and that's why I don't like looking at the samples. I think you always have to go back to the criteria and if you go back to the criteria, it's really a B. Really, even though the top section is definitely an A, they haven't finished ... they haven't finished this ...

It may be that the use of exemplars to demonstrate the qualities of each of the standards (A-E) could more accurately describe an ideal example or the typicality of the standard than work samples that partially reflect the qualities and do not comprehensively depict the typical instance of the standard.

Ecological validity

From the interview data it was apparent that teachers were 'caught' by the nature of the tasks. Standardisation brings issues 'out of context' so that teachers feel compelled to teach to the task which raises issues of ecological validity (Black and Wiliam, 1998). In the following excerpt from the conversation data a teacher explains how the task of report writing was not familiar to her students.

F2: My kids didn't do well in that because it wasn't something that I had familiarised them with and because of the way it was set out. It was just an unusual sort of setting out that they hadn't even encountered before ...

F1: And do grade 3, it's a lot of recounts ... do they do much of this report writing? ... because I remember the information report as a big thing.

F2: They do information reports as one of their main genre but not report writing like this.

F1: No

F2: It's mostly book work.

F3: In year 3 it's mostly recounts, in year 4 it's report writing, information report writing.

This indicates an issue related to the alignment of the curriculum with the assessment task. As the Essential Learnings are being developed and implemented teachers are adjusting their teaching programmes to meet these new demands.

Specificity of Criteria

Difficulties arose for teachers moving from the specifics of the criteria within a task to the broader evaluations of the standards as used in the KLAs. The language of the standards in terms of the clarity, the detail provided and the notions of 'quantity' caused a dilemma for some, yet was welcomed by others. For example, the quantitative wording of the standards (see the task specific descriptors for products one and two in Appendix Two) proved an issue.

F2: Yeah, identified more than one key detail.

M2: Identified more than one detail.

M1: They actually only put down one key detail.

F1: I did too actually. I said only one relevant detail.

M1: That's what I thought – it was sort of ...

F2: She identified the 8 year old boy named Jay and he tripped over a rock and he was badly injured.

F1: But she didn't answer any of the other things.

M1: That's all one detail.

F2: No but the criteria descriptor says 'more than one detail'. Have a look.

M1: But I thought that was all just ... one detail.

F1: But was it one detail that was relevant to the criteria? That's what I marked when there was one detail. Then she goes into what ... but that's not what the question asks. That's quite a different question.

In considering this point it is important to heed Sadler's (2007: 390) concern about the perils of meticulous specification of assessment criteria. He has indicated that in the context of post-compulsory schooling "...assessment criteria have been reduced to pea-sized bits to be swallowed one at a time – and for each bit, once only." He goes on to

suggest that it is "... an extreme interpretation of what criterion-reference assessment implies" that has accentuated the trend to coaching. We would argue that it is important in the context of assessment reform that these insights are considered to reduce the likelihood of 'fine-grained prescription', 'atomised assessment' (Sadler, 2007) and "that achievement can come to be seen as little more than criteria compliance in pursuit of grades" (Torrance, 2007: 286). If these issues are to be addressed teachers will need to receive assistance. However, the standards were viewed as valuable for understanding year level requirements as well as positioning students within year levels.

Discussion

In the discussion of the emergent issues that follows we have used Wenger's (1998) components of a social theory of learning: 'meaning', 'practice', 'identity' and 'community' in the specific context of a developing 'community of practice' in social moderation.

Meaning

The issues that have emerged from this early phase of study have shown that there is a vast range of understandings of moderation processes and procedures. This is to be expected given that the teachers are from different social and cultural contexts with varying histories of engagement in moderation practice. Teachers are reasonably consistent in their accounts of the purposes of moderation but still present varied understandings. Some teachers described the aims of moderation, its practice and outcomes consistently with the descriptions offered in the QSA guidelines. Some teachers appear to understand many of the features of moderation.

A minority of teachers' responses turned the focus of moderation from looking at the quality of student work to the quality of the judgement. For instance from the interview data:

... it's almost like an outside body um, keeping the teachers on the same page and that could be a good thing ...

...we thought it was a B and others on the table were saying it was a C- what is the right grade?

This notion of a piece of work having a 'right' (intrinsic) grade is common amongst not only teachers but also examiners. The problem is that this belief can become a 'truism' in that we can search and find the right grade for the piece of work and the student but this can never be as the grade is the property of the student, the task and the person making the judgement.

Such responses are important to consider as the teacher's assessment practice becomes the subject of focus and the perceived evaluative nature of moderation may become threatening to teachers in terms of their own identity formation as they become a member of this community of practice as a 'legitimate peripheral participant' (Lave and Wenger, 2003).

Practice

The majority of teachers interviewed valued the moderation process as a way to view and share pedagogical practices. However, despite the perceived rewards from teachers' participation, the time required for organisation and actual involvement was considered to be a constraint. Teachers identified the need for further practice and time to feel comfortable about sharing their own judgements. This can be understood from a social theory of learning in terms of 'learning as doing' and 'learning as belonging' to a community. As one interviewee expressed this:

Time needs to be given to the teachers to feel comfortable in 'sharing' work with a stranger. It needs to be stressed that it's for the purpose of the children to see if the work is of a consistent standard across schools.

Identity

An environment that was non-threatening, non-judgemental and where teachers respected each other's opinions was reported to be important for moderation to be most effective. From a social theory of learning perspective the formation of such a community requires participants in the social configurations of moderation practice to engage in further practice before competence as a moderator becomes recognisable to the community and a sense of becoming or identity is more fully formed. While it is really too early to make any claims about the formation of identity there was some evidence that a notion of identity was emerging for some. For example from the interview data:

... I was actually spot on with everybody else, so that was just really good for my own knowledge and to know that you can be confident in making your own judgements.

... all along I have been considered a hard ... you know taskmaster and hard marker by a lot of other teachers, I feel like I've been on the right track, so that was good.

Community

Different methods of intra-school moderation, some appearing to be more effective than others have emerged with the introduction of the QCAR framework. In one example from the analysis of interview data, all of the QCATs were marked by the teachers together.

... we really did it one by one. We didn't just mark one and like go for it. We talked through with each other like 'Oh, I don't really know what this is' then talked it out and justified. ... It took us a whole day. ... I don't think we were good enough to look at the standards and say 'they have this', the sample and the discussion was much better.

Again from an analysis of the interview data another model of moderation involved one teacher waiting until others in the school had marked their tasks before proceeding to mark her own entire cohort. Yet another procedure adopted for moderation was a distribution of tasks amongst teachers to be assessed and then moderated. First teachers marked products one and two, together, and then marked products three and four on their own. When the teachers came together again and collated their final results, they found there was consistency of standards. The final method of moderation that emerged was for teachers to mark their own class' tasks and then to remark the work without names.

There were complex issues arising from these methods of moderation. Although teachers were concerned with the time taken by moderation, they were keen to moderate every piece of work. Moderation as a process was valued by some teachers to the extent that time issues were put aside for this process to occur. Generally it was found that the attitude teachers approached moderation determined what value they gained from the process. There is a requirement that teachers are informed of the moderation process and the expectations of moderation procedures and practices before participating in such events if maximum benefits for their own professional development are to occur. There was some indication that involvement in moderation can impact positively on teachers' pedagogical practices.

The teachers were focused on improving their teaching and their assessment practices but there was little if any mention of assessment to be used for the development of student learning or the involvement of the students in this process. These initial findings have indicated that there is a need to support teachers in their development of assessment skills and assessment literacy. There is a vast range of experience and understandings of the moderation process, using standards to inform judgement practices, and the application of this knowledge to the teaching/learning process within middle school teachers. Overwhelmingly though, social moderation is viewed by teachers as a worthwhile professional process in which to be involved.

Implications for teacher development

Building teacher assessment capacity is the major challenge that requires provision of systematic teacher support. A study conducted by Klenowski (2006) involving teachers in the use of standards for moderation purposes also found gaps in teachers' subject knowledge, assessment practice and/or pedagogical understanding. Research evidence illustrates how teachers who are trained to mark common assessment tasks find the moderation process richly rewarding (Reid, 2003; Klenowski, 2006, 2007). Identified benefits include; teachers' improved understanding of the standards expected at the particular level at which the task is set and possible gaps in teachers' own teaching, learning programmes and curriculum priorities. Teachers welcome such opportunities to discuss student work.

Professional learning communities need to be established at the local level to provide teachers with the support required in developing skills in assessment. Financial and human resources in the form of allowing teacher release is required so that teachers can attend related professional development opportunities and moderation meetings. For teachers, particularly those located in remote settings, the development of online communities through moderation meetings may be an avenue in which these challenges are met. In a developmental context it is important that teachers' needs are met by support that parallels the pilot and trialing phases of the framework. It has been shown (Black, Harrison, Lee, Marshall, & Wiliam, 2003; Klenowski, 2007) that teachers respond to principles and practices that they can relate to and that are grounded in their own contexts. They do not change their practices based on research or evaluation evidence alone; they need examples of implementation of assessment practices and procedures by teachers with whom they can relate.

Teachers' participation in various stages of moderation is one of the most important contributors to developing teachers' confidence in applying standards for grading students' assessment tasks as well as professional skills enhancement (Klenowski, 2007). When the scope of moderation is sufficiently wide to situate assessment evidence (student-generated artefacts) in relation to issues of task design, teachers are afforded opportunities to probe the vital links between curriculum, pedagogy and the resultant quality of student achievement. Ironically, in Australia, recent debates regarding the need for education reform and greater consistency in education in all states and territories (Reid, 2005, 2006; Masters, 2006; Smyth, 2006) have been driven by the Australian Federal Government's push for: national standards; reporting nationally comparable data on student achievements and reporting to parents and the wider community on schooling outcomes for accountability purposes. The federal government's interest is in ensuring that schools maintain the country's international competitiveness. In these debates, moderation is not a priority which indicates a policy drive to locate standards within system measurement mechanisms, and away from the practice of classroom teachers. It is this market shift of locus of control towards the use of standards for reporting and accountability purposes that has impacted on developments in all Australian states including Queensland. Given this changed policy context the professional development implications for teachers are profound.

Conclusion

We have identified the gap in the research regarding moderation and the role of teacher judgement. We have designed research to fill this gap and believe there are several ways forward. The first way forward acknowledges the context of the knowledge society (Drucker, 1969; UNESCO, 2005; Schleicher, 2006). Demands for greater sharing of information and knowledge that is easily retrievable to the greatest number of individuals and enterprises requires that the potential for moderation is researched online in a virtual context, as distinct to face-to-face meetings. Digital materials such as student portfolios with matching statements of standards and commentaries on the justification for the grade that demonstrates work at different standards will be far more accessible to teachers throughout the state. The high level of insecurity around the standard could be addressed by exemplification using digital means. In providing a digital match between the requirements for the standard and the evidence in the portfolio it is important to understand how compensations have occurred. For example, the trade offs, the strengths and limitations in the performance evidence contained in the portfolio could be illustrated and in this way some of the threats to validity in the assessment of the portfolio, such as construct under representation and construct irrelevant variance (Messick, 1989) could be addressed.

In line with our theoretical framework that values socio-cultural and situated perspectives of learning we agree with Elwood, (2006: 228) that in assessing students' achievements we need to consider 'what they bring to the classroom setting as a consequence of their participation in a myriad of other cultural contexts'. Insider insights of teachers about how standards apply and how judgements can accommodate not only knowledge of the standards but also knowledge that enables the assessor to be responsive to school

contextual information is an area of particular interest given that there are greater numbers of Aboriginal and Torres Strait Islander students in the North of Queensland. Related to this concern about context for equity purposes it is equally important to consider the reliability issues associated with the conditions in which the tasks are completed and assessed. If moderation activities are to be incorporated into the professional practices of middle school teachers, we need to provide a better understanding through research that incorporates case studies of how to document the conditions under which assessments are undertaken and to make explicit how this knowledge is used in conjunction with knowledge of stated standards. In so doing we cannot ignore the critical access questions of - Who gets taught and by whom? Whether there are differences in the resources available? What is incorporated from the cultures attending? (Stobart, 2005) We are hopeful that this ARC funded project may provide further insights to improve our understanding of teacher judgement and its role in an assessment-driven accountability context.

References

- Arnold, P. & Smith, J. D. (2003) Adding connectivity and losing context with ICT: contrasting learning situations from a community of practice perspective, in: M. Huysman, E. Wenger & V. Wulf (Eds) *Communities and technologies* (The Netherlands, Kluwer Academic Publishers), 465-484.
- Assessment Reform Group. (2002) *Testing, motivation and learning* (Cambridge, UK, University of Cambridge Department of Education).
- Black, P. and Wiliam, D. (1998) Assessment and classroom learning, *Assessment in Education*, 5(1), 7-74.
- Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2003) *Assessment for Learning: putting it into practice* (Maidenhead, Open University Press).
- Black, P. & Wiliam, D. (2005) Lessons from around the world: how policies, politics and cultures constrain and afford assessment practices, *The Curriculum Journal*, 16(2), 249-261.
- Broadfoot, P. & Black, P. (2004) The first ten years of assessment in education, *Assessment in Education: Principles, Policy & Practice*, 11(1), 7-26.
- Cumming, J.J., Wyatt-Smith, C.M., Elkins, J., & Neville, M. (2006) *Teacher judgment: building an evidentiary base for quality literacy and numeracy education* (Brisbane, Qld, Queensland Studies Authority).
- Department of Education and the Arts (DETA), (2005) *Queensland Curriculum, Assessment and Reporting Framework* (Brisbane, Australia, Department of Education and the Arts, Queensland Government).
- Drucker, P. (1969) *The age of discontinuity. Guidelines to our changing society* (New York, Harper and Row).
- Education Queensland, (2002) *The Report of the Assessment and Reporting Taskforce* (Brisbane, The State of Queensland Department of Education).
- Elwood, J. (2006) Formative assessment: possibilities, boundaries and limitations, *Assessment in Education: Principles, Policy and Practice*, 13(2), 215-232.
- Elwood, J. & Klenowski, V. (2002) Creating communities of shared practice: assessment use in learning and teaching, *Assessment and Evaluation in Higher Education*, 27(3), 243-256.

- Goodyear, P., Banks, S., Hodgson, V. & McConnell, D. (Eds) (2004) *Advances in research on networked learning* (Boston, U.S., Kluwer Academic Publishers).
- Hara, N. & Kling, R. (2000) Students' distress with a web-based distance education course. An ethnographic study of participants' experiences, *Information, Communication and Society*, 3(4), 557-579.
- Harlen, W. (2005a) Teachers' summative practices and assessment for learning – tensions and synergies, *The Curriculum Journal*, 16(2), 207-223.
- Harlen, W. (2005b) Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes, *Research Papers in Education*, 20(3), 245-270.
- Harlen, W. (1994) Concepts of quality in student assessment. Paper presented at Annual Meeting of the American Educational Research Association, April 4-8, New Orleans.
- Hine, C. (2000) *Virtual Ethnography* (London, UK, Sage).
- Hutchinson, C. & Hayward, L. (2005) The journey so far: assessment for learning in Scotland, *The Curriculum Journal*, 16(2), 225-248.
- Ingvarson, L. (1990) *Enhancing professional skill and accountability in the assessment of student learning*. Paper presented at the annual Meeting of the American Educational Research Association, Boston, USA.
- James, M. (2006) Assessment, teaching and theories of learning. In J. Gardner (Ed) *Assessment and Learning* (London, UK, Sage Publications).
- Klenowski, V. (2007) *Evaluation of the effectiveness of the consensus-based standards validation process* (Brisbane, Qld, Department of Education, Training and the Arts). Available online at: http://education.qld.gov.au/corporate/newbasics/html/lce_eval.html (accessed 18 October 2006).
- Klenowski, V., Askew, S. & Carnell, E. (2006) Portfolios for learning, assessment and professional development in higher education, *Assessment and Evaluation in Higher Education*, 31(3), 267-286
- Klenowski, V. (2006) *Evaluation report of the pilot of the 2005 Queensland Assessment Task (QAT)* (Brisbane, Qld, QSA). Available online at: <http://www.qsa.qld.edu.au/research/reports.html> (accessed 18 October 2006).
- Klenowski, V. (2002) *Developing portfolios for learning and assessment: processes and principles* (London, RoutledgeFalmer).
- Lave, J. & Wenger, E. (2003) *Situated learning: legitimate peripheral participation* (New York, USA, Cambridge University Press).
- Linn, R. L. (1993) Linking results of distinct assessments, *Applied Measurement in Education*, 6(1), 83-102.
- Luke, A. (2003) After the marketplace: evidence, social science and educational research, *Australian Educational Researcher*, 30(2), 87-107.
- Malone, L., Long, K. & De Lucchi, L. (2004) All things in moderation, *Science and Children*, 41(5), 30-34.
- Masters, G. (2006) National curriculum collaboration in senior secondary schooling, *Curriculum Perspectives*, 26(3), 56-60.
- Matters, G. (2006) *Statistical moderation and social moderation around Australia*, Paper presented at the International Association for Educational Assessment Conference, May 2006, Singapore.

- Maxwell, G. (2002) *Moderation of teacher judgements in student assessment (Discussion Paper)* (Queensland School Curriculum Council, Brisbane).
- Messick, S. (1989) Validity, in: R. L. Linn (Ed), *Educational Measurement* (3rd Ed) (New York, Macmillan), 13-103.
- Queensland Studies Authority. (2007a) *Information statement February 2007: Essential learnings draft 2* (Brisbane, Australia, QSA).
- Queensland Studies Authority. (2007b) *Information statement April 2007: Standards draft 2* (Brisbane, Australia, QSA).
- Queensland Studies Authority. (2007c) *September 2007 Trial common assessment task: Teacher guidelines* (Brisbane, Australia, QSA).
- Queensland Studies Authority (2007d) *Year 4 English trial statewide assessment teacher guidelines* (Brisbane, Australia, QSA).
- Reid, L. (2003) 'Now I know what to do!' *Teachers talking about writing assessment*, Paper presented at the Scottish Educational Research Association Annual Conference, Perth, Scotland, 27-29 November. Available online at: <http://www.leeds.ac.uk/educol/documents/00003486.htm> (accessed 18 October 2006)
- Reid, A. (2005) Rethinking national collaboration: towards an Australian curriculum. Available online at: http://www.dest.gov.au/sectors/school_education/publications_resources/profiles/rethinking_national_curriculum_collaboration.htm (accessed 16 June 2006).
- Reid, A. (2006) Is consistency the only rationale for national curriculum collaboration? *Curriculum Perspectives*, 26(3), 66-70.
- Sadler, R. (2007) Perils in the meticulous specification of goals and assessment criteria, *Assessment in Education: Principles, Policy and Practice*, 14(3), 387-392.
- Sadler, R. (2005) Interpretations of criteria-based assessment and grading in higher education, *Assessment and Evaluation in Higher Education*, 30(2), 175-194.
- Schleicher, A. (2006) *The economics of knowledge: why education is key for Europe's success* (Brussels, The Lisbon Council).
- Slack, J.D. & Wise, J.M. (2005) *Culture and technology: a primer* (New York, US, Peter Lang Publishing).
- Smyth, J. (2006) The politics of reform of teachers' work and the consequences for schools: some implications for teacher education, *Asia-Pacific Journal of Teacher Education*, 34(3), 301-319.
- Stehlik, T. & Carden, P. (Eds) (2005) *Beyond communities of practice: theory as experience* (Flaxton, Australia, Post Pressed).
- Stobart, G. (2005) Fairness in multicultural assessment systems, *Assessment in Education: Principles, Policy and Practice*, 12(3), 275-287.
- Torrance, H. (2007) Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning, *Assessment in Education: Principles, Policy and Practice*, 14(3), 281-294.
- United Nations Educational Scientific and Cultural Organisation (2005) *Towards knowledge societies*, UNESCO World Report, (Paris, UNESCO). Available online at: <http://www.unesco.org/publications> (accessed 16 October 2006).
- Wenger, E. (1998) *Communities of practice: learning, meaning and identity* (Cambridge, U.K., Cambridge University Press).

- Wilson, M. (Ed) (2004) *Towards coherence between classroom assessment and accountability*, The 103rd Yearbook of the National Society for the Study of Education Part 2, (Chicago, Chicago University Press).
- Wyatt-Smith, C, (2005) Examining how teachers judge student writing: an Australian case study, *Journal of Curriculum Studies*, 37(2), 131-54.
- Wyatt-Smith, C., Cumming, J. & Elkins, J. (2005) Redesigning assessment, in D. Pendergast & N. Bahr (Eds), *Teaching middle years: rethinking pedagogy and assessment* (Crows Nest, Australia, Allen & Unwin).
- Wyatt-Smith, C. & Castleton, G (2004) Factors affecting writing achievement: mapping teacher beliefs, *English in Education*, 38(1), 37-61.
- Wyatt-Smith, C., Castleton, G., Freebody, P., & Cooksey, R. (2003) Teachers' qualitative judgements: a matter of context and salience. Part I. *Australian Journal of Language and Literacy*, 26(2), 11-32.

Appendix One

Project Overview

Investigating standards-driven reform in assessment in the middle years of schooling

Project Aims

- **Examine** how teachers use stated standards to judge quality of student work;
- **Develop, trial, and evaluate models** of moderation to support consistency of teacher judgement and to inform standards-based reporting to students, parents/ carers, and systems;
- **Assess** the utility, including cost-effectiveness, of **models** of moderation for their responsiveness to different pedagogical and geographical contexts and student diversity; and
- **Produce a best practice framework** for teacher judgement in statewide moderation in years 1 to 10 including exemplars of how standards are met.

Research Questions

The project's key research questions are addressed in three related inquiries:

1. How do stated standards work to inform and regulate judgement in different curriculum domains? (*Inquiry 1*)
2. What are the configural properties of teacher judgements about the standard of the work assessed and how are they explicitly shared between teachers? (*Inquiry 1*)
3. Does the social practice of moderation involving the application of explicitly-defined standards result in changed judgements about students' work? (*Inquiry 2*)
4. Does moderation using standards result in consistency of teacher judgement? (*Inquiry 3*)

Inquiries

Inquiry 1 – Judgements of different assessment evidence (survey)

How teachers develop, share, and apply knowledge of standards-referenced assessment to two different types of assessment evidence (CATs, LATs).

Inquiry 2 - Development, Trial and Evaluation of Moderation Modes and Practices

1. Develop, investigate and compare the application of three modes of moderation:
 - face-to-face,
 - a combination of face-to-face and ICT-mediated,
 - and ICT-mediated contact where communication is exclusively online.

Different ways of facilitating moderation to be compared including various mixes of pre-meeting perusal of folios; individual, pair, or trio perusal of folios; whole-school folio moderation; by-year level moderation; and novice-expert groupings - both within school, and between school moderation practices.

2. Focus on role of principals, deputy-principals, or other school curriculum leaders in intra- and inter-school moderation.

3. Focus on how teachers participating in moderation discourses negotiate and ascribe meaning to stated standards and arrive at judgements of student work. Multiple outcome measures to be used to evaluate models of moderation and teachers' use of standards as they relate to educational inclusion policy and practices, with specific reference to Disability Standards for Education.
4. Teachers' conceptions of standards-referenced judgements and their actual judgements will be documented both pre-and post-moderation on five occasions throughout the study investigating changes in actual judgements of quality that result from moderation and the shared discussion of how standards apply. Also changes that may occur in judgement practices themselves.

Inquiry 3 Portability of standards across sites and learning domains

1. Documents teachers' decision-making processes in applying standards (e.g. A to E).
2. How teachers apply the standards at both mid-standard and at standard-entry giving exemplars illustrating the standards.
3. Teachers to use standards to moderate student work from schools outside their immediate geographic location - a blind-review of the work, the moderation task being to judge its quality against the stated standards (i.e. moderating the work of students unknown to them).

Summary of Data Collection Instruments

- **Survey**
- Audio **recording** of moderation meetings; and/or audio & screen capture recording of online moderation sessions
- Semi-structured **interviews** with focus teachers – pre and post moderation

Participants

Yrs 4, 6, 9 (English, Maths and Science)

Year Level	N: Teachers	N: Focus Teachers
Yr 4	Approx 80	20
Yr 6	Approx 80	20
Yr 9	Approx 80	20

Representative of one or more of the following categories: a) remote and rural; b) metropolitan; c) Indigenous and Torres Strait Islanders; d) cultural and linguistic diversity; e) low socio-economic; f) distance education; g) disabilities; h) high transience; i) cross sector involvement of Catholic, Independent and State schools, and j) gifted and talented.

Phases of the research:

1. **March 2007 – June 07** Reference group established, ethics clearances obtained, schools selected and invited; survey design and distribution, contact and begin familiarisation with Web Ex (on-line screen capture). Commence data collection of moderation meetings including pilot of audio equipment.
2. **July 2007 – December 2007** Moderation meetings re CATs / LATs - teachers use standards to judge students' achievement – 240 teachers will use standards to judge

students' achievement in Yrs 4, 6, 9 in English, Maths and Science CATs. Data will be collected involving 60 focus teachers including using audio recording and/or audio/screen capture recording of online moderation sessions of teachers' focused discussions on student samples of work as they use stated standards to analyse the characteristics of the work generated for CATs; this data collection to be replicated for the LATs; teacher interviews – pre & post moderation.

3. **January – December 2008** Portability of standards across sites and blind review, produce best practice framework for teacher judgement in state-wide moderation Years 1-10 including exemplars of how standards are met.
4. **January 2009 – December 2009** Final forum, implications of study for state-wide implementation of standards; preparation and submission of final report. Publication and presentation of papers reporting of findings of the study will be undertaken during phases 1-4 and a final evidential base of exemplars and models of moderation will be provided for the industry partner in phase 4 synthesising the 3 inquiries to include policy recommendations.

Appendix Two

Guide to making judgements - Year 4 English

Student

Overall grade A B C D E N

D i m e n s i o n s K n o w i n g & U n d e r s t a n d i n g W a y s o f W o r k i n g	Assessable Elements	Task-Specific Descriptors				
		A	B	C	D	E
		Understands main ideas within report texts <i>Product 1</i>	Accurately identifies the main idea common to both texts and all key details including potential intent.	Identifies the main idea to both texts and most key details (will not include both time and date).	Identifies the main idea to both texts and key details of when (date or time), who (one person only), how and why.	Identifies some key details.
	Analyses and compares text purpose, type, audience and subject matter in report genres. <i>Product 2</i>	Accurately identifies all texts' similarities/ differences in detail (particularly purpose and intent).	Accurately identifies most texts' similarities/ differences in some detail, including either purpose or intent.	Identifies most texts' similarities/ differences in some detail.	Identifies at least two similarities/ Differences.	Suggests one similarity/ difference.
	Writes and shapes a written report to suit purpose, subject matter and audience <i>Product 3</i>	Constructs a cohesive, informative, logically organised report. Includes all key organisation-al inform-tion (who, why, where and when). Extensive detail is given about the event in appropriate language.	Constructs an informative, sequenced report. Includes key organisation-al information. Some detail is given about the event in appropriate language.	Uses some features of report genre, giving opinions rather than facts. Includes some key organisation information.	Attempts to construct a report by listing basic information.	Attempts to describe an event.
	Understands appropriate simple sentence structure, spelling and punctuation strategies. <i>Product 3</i>	Uses some paragraphing and varied sentence structure. Few spelling and punctuation errors.	Uses one paragraph and some variation in sentence structure. Few spelling and punctuation errors.	Uses one paragraph and simple sentence structure. Spelling/ punctuation errors do not impede meaning.	Uses a brief series of individual sentences and fragments. Spelling/ punctuation errors do not impede meaning.	Sentence structure and spelling/ Punctuation errors impede meaning.
	Reflects on literary choices made to portray people, places and events in a written report. <i>Product 4</i>	Makes detailed reflective comments on words chosen and fair reporting.	Makes brief reflective comments on words chosen and fair reporting.	Makes a general comment on words chosen and fair reporting.	Makes a general comment on words chosen or fair reporting.	Makes a comment about own reporting.