Examining how moderation is enacted within an assessment policy reform initiative: You just have to learn how to see

Claire Wyatt-Smith and Val Klenowski

Claire Wyatt-Smith is Professor of Education and Dean of the Faculty of Education at Griffith University. She has been chief investigator on a number of Australian Research Council projects and other funded research studies focusing on teacher judgment, standards and evaluative frameworks as these apply in policy and practice. Current research includes an investigation of the dependability of teacher assessment for statewide reporting in Queensland. A current ARC project includes a longitudinal study of standards-driven assessment reform in middle schooling, the focus being on teacher judgment and moderation, both face-to-face and ICT mediated.

Professor Claire Wyatt-Smith
Dean
Faculty of Education
Mt Gravatt campus
Griffith University
Brisbane
Queensland 4111
AUSTRALIA
Tel: +61 (07) 373 55647
Fax + 61 (07) 373 56802

Email: c.wyatt-smith@griffith.edu.au

Val Klenowski is Professor of Education at the Queensland University of Technology and is currently Chief Investigator on three Australian Research Council Linkage projects that focus on assessment principles, policy and practice: digital portfolios for marginalised youth, culturally-fair assessment for Indigenous students and moderation practices using a combination of ICT mediated and face-to-face meetings. She has researched and published in the fields of curriculum, evaluation, assessment and learning. She currently leads and co-ordinates the largest Professional Doctoral Program in Australia.

Professor Val Klenowski
School of Learning and Professional Studies
Queensland University of Technology
Victoria Park Road
Kelvin Grove
Queensland 4059
AUSTRALIA
Tel: +61 (07) 3138 3415
Fax: (07) 3138 3987

Email: val.klenowski@qut.edu.au

Examining how moderation is enacted within an assessment policy reform initiative: You just have to learn how to see

Claire Wyatt-Smith and Val Klenowski

Introduction

In 2008, the new Labor Government in Australia has introduced plans for a National Curriculum in mathematics, science, history and English in primary and secondary schools to be implemented in 2011 and extended to involve geography and languages other than English in a second phase. The intention is to establish a standards-referenced framework to "invigorate a national effort to improve student learning in the selected subjects" (National Curriculum Board, 2008: 3). In 2007 states and territories in Australia developed individual approaches to the use of standards in the implementation of curriculum, assessment and reporting. The latter involves schools to report using A-E grades that is consistent with the Federal government's requirement. This paper reports on a four year Australian Research Council Linkage[i] project, being conducted in the Australian State of Queensland. The major industry partner for this project is the Queensland Studies Authority (QSA). The intent of such partnerships is for research, conducted by academics in liaison with policy officers, to inform policy development. Too often policy officers have to 'grasp the complex remit quickly and take action' (Saunders, 2005) without the benefits of policy-related research. The findings from this study are being used developmentally to inform and influence the policy environment, particular assessment initiatives and practices.

Background

For the first time in Queensland, teachers in the middle years of schooling (Years 4 to 9) are using defined standards to form judgements of student work and are engaging in social moderation. The focus of this study is on the use of state standards to promote consistency of teacher judgement of student work. A central research question is: How do teachers in the middle years of schooling, use stated standards in moderation to achieve consistency of judgement? In the years of schooling from Pre-school to Year 10, teachers have been required to use stated curriculum outcomes written as developmental markers, with a primary focus on their application to teaching and learning. They have not been required to use standards for assessing and grading purposes, nor has there been a requirement for them to undertake inter- or intra-school moderation as part of system efforts to support consistency of teacher judgement. The conceptual leap expected of teachers requires them now to assess student achievement on centrally-developed assessment tasks (Queensland Comparable Assessment Tasks or QCATs) using five defined standards (A to E), and to achieve consistency of judgement and reporting using standards (see Table 1). The Teaching and Learning Division of QSA has had responsibility for devising and developing the Queensland Curriculum, Assessment and Reporting Framework (QCAR) with an emphasis on products (QCATs, the guide to making judgements, annotated samples of student work, database of assessment tasks) rather than processes. The research reported in this paper involves a collaborative approach to policy reform by investigating policy development and enactment in the context of a trial or pilot study before full implementation. The interactive nature of this reform involves research at the local professional level of districts and their schools.

Table 1. The QCAR Framework Retrieved from QSA website (http://www.qsa.qld.edu.au/assessment/qcar.html) on 25 March 2008.

Essential Learnings	To identify what should be taught (key knowledge, facts, procedures and ways of working) and what is important for students to have opportunities to know, understand and be able to do.
Standards	To provide a common frame of reference and a shared language to describe student achievement.
Online Assessment bank	To support everyday assessment practices of teacher through access to a range of quality assessment tools.
Queensland Comparable Assessment Tasks (QCATs)	To provide information on what students know, understand and can do, in a selection of Essential

	Learnings. QCATs are intended to promote consistency of teacher judgements across the state.
Guidelines for reporting	To support consistency of reporting across the state.

Attaining coherence between classroom assessment and system level accountability, that includes system interests in transparency of schooling outcomes, has been debated (Frederiksen & White, 2004; Wilson, 2004), the centrality of teachers' judgement practice in achieving such coherence is the focus of this research (Wyatt-Smith, Klenowski and Gunn, in press; Klenowski and Adie, in press).

The Study

The research questions for this project include:

- How do stated standards work to inform and regulate judgement in different curriculum domains?
- What processes including social interactions do teachers rely on to inform their judgement decisions?
- What are the properties or characteristics of teacher judgements and how are these (as distinct from outcomes of grading decisions) shared or made available to other teachers?
- Does the social practice of moderation involving the application of explicitly-defined standards result in changed judgements about students' work?
- Does moderation using standards result in consistency of teacher judgement?

Both quantitative and qualitative methodologies are used and the data collection methods include focus teacher pre- and post-moderation interviews and conversation data of the moderation meetings. All interview and meeting data are progressively transcribed in full for analysis. A multi-theoretical approach to data analysis has been adopted. Each data set has been analysed by more than one researcher drawing on the following theoretical orientations. Sadler's (1989) writing on the formulation and promulgation of standards, specifically that standards written as verbal descriptors are necessarily fuzzy (as distinct from sharp, such as numeric cut-offs). Second, are understandings of learning theory in which assessment is central to the learning and teaching cycle (Shepard, 2000). Third, reflecting the socio-cultural framing of the project, are understandings about language itself as inherently social and cultural (Kress, 2000: 401). In this paper we have adopted Kress' notion that the teacher or sign-maker is "constantly transformative of the set of resources of the group and of her/himself." 'You just have to learn how to see' applies to teachers when they use standards to assess student work to achieve consistency in judgement during moderation. The analysis of talk data captured during these meetings together with the analysis of the interview data reveal that there is interaction with several modes of representation and communication. As in Kress' analysis we have identified the process to be multimodal and have analysed how during the interactions there is a specialisation of representational and communicational modes.

To date the study (Wyatt-Smith et. al. in press) has identified an empirically derived framework of categories of referents that teachers have used to confirm or challenge proposed grades during the moderation meetings. These two categories of referents include, first, the textual materials provided by QSA (i.e. the products such as the guide, and annotated student work samples), together with the readings that the teachers have made of these; second, types of tacit and social knowledge evident in the recorded talk but missing from the official documentation. These include discipline (subject) knowledge, knowledge of the official curriculum; the teachers' prior evaluative experience; knowledge of individual students, and knowledge of what the 'average' student could reasonably be expected to demonstrate at a given year level.

Policy Context

QSA's P-12 draft assessment policy states that for social moderation to work effectively, there are six suggested requirements. One that is pertinent to our discussion is that there are: "syllabuses that clearly describe the standards of learning and standards of assessment" and the second requirement that is relevant is that there are "teacher discussions of the quality of the assessment instruments and the standards of student work" (QSA, 2008: 1). It should be stated that this draft policy emerged in April

2008 from the Student Achievement Division of the organization concurrent with the QCAR initiative, a responsibility of the Teaching and Learning Division.

The QCATs have been designed and are being trialed and developed by the QCAR team using classroom teacher input and feedback. The tasks have been written to provide some opportunity for 'authentic assessment' as opposed to more paper and pencil test formats. Due to the pressures of timelines, budgets and product expectations the alignment of the teaching of the Essential Learnings (ELs) with the constructs of the QCATs has not always been possible because the ELs are being introduced at the same time as the tasks are being trialed, developed and administered. Thus some ELs have not been taught, yet some teachers have had to administer tasks to their students who have not had the opportunity to learn the underlying constructs.

This has raised issues related to validity and is apparent in the analysis of the interview and moderation meeting data. Identifying the key constructs for teaching and learning and subsequent assessment is fundamental to the concept of validity. All assessments are based on a sample of behaviour or performance in which we are interested and it is from the sample that we generalise to 'the universe of that behaviour'. The 'fidelity of the inference drawn from the responses to the assessment is what is called the validity of the assessment' (Nuttall, 1987: 110-111). This is why the specification of the domain of behaviour in which we are interested is critically important.

Crooks, Kane and Cohen (1996) identified the threats to validity that developers of assessment tasks should address. For example, the following threats linked to the scoring or grading of the student's performances on a task are identified as: the scoring or grading fails to capture important qualities of task performance, there is undue emphasis on some criteria, forms or styles of response, there is a lack of intra-rater and/or inter-rater consistency and the scoring or grading is either too analytic or too holistic. Further threats to validity include construct representation and construct variance. These threats suggest that teachers need to be aware of the key constructs to facilitate their judgement practice. Construct representation refers to the extent to which the task samples the knowledge, skills and/or constructs it is intended to assess. When "the test [or task] is too narrow and fails to include important dimensions or facets of the construct" there is construct under-representation (Messick, 1989: 34). Construct irrelevance refers to the construction of the task and the reliability of the results (Messick, 1989). Construct-irrelevant variance exists when the "test contains excess reliable variance that is irrelevant to the interpreted construct" (Messick, 1989, p.34). This form of construct-irrelevant variance is regarded as a contaminant with respect to the score or grade interpretation.

Discussion and Findings

What follows is an analysis of data that illustrates how teachers are 'learning to see' from their interpretation of the standards based on a careful reading of textual materials. During the interactions of the moderation meetings there are expectations expressed in the pre-moderation interviews of verification, 'vindication' and validation of judgements. There is also evidence of the teachers being transformative of the set of resources of the moderation meeting and him/herself. The modes of representation and communication are speech used as a mode of commentary, critique and ratification.

Standards Informing Judgement

In the following pre-moderation interview Chris indicates what he hopes to gain from the moderation meeting, he describes how he might 'learn to see' the standard as he has interpreted it. The transformative nature of the available resources become apparent:

> Chris: I think I'll be interested to see, I'm, the, the most important part for me will be, will be meeting the teachers and hearing what they say, and seeing also, for my own thing, seeing if, I mean, I'll go along there with a pre-conceived notion already of what those, the QCATs and thinking, "Well, okay, I think so-and-so's going to be an A," so in the back of my mind I'll have their work. I'm really interested to see whether that matches up with what other teachers think about it, too. And if it does it vindicates me. If it doesn't I have to go back and say, "Okay, I am marking too hard, I am marking too easily." So that will be good, that will be really good.

The importance of the interaction during the moderation meeting appears to be acknowledged by this teacher as he reveals how his tacit knowledge of the student and his expectations will either be confirmed or challenged. Knowledge of the student, specifically as it relates to affective dimensions (e.g. dispositions and attitudes), can also act as a threat to validity. Further, knowledge of the student's work that has been previously assessed is deemed irrelevant. In the grading of the student's work on QCATs the focus should be on the qualities of the completed work in relation to the constructs being assessed. The interactions are important in the transformation that is anticipated by this teacher through the resources of the group and the teacher's subjectivity. His interest in aligning his judgement with that of other teachers is apparent when he states that in matching up his judgement with those of other teachers the outcome will 'vindicate' *him*.

In what follows a teacher identifies the criteria (assessable elements) related to the science QCAT, and the task itself, as problematic in informing judgement. The teachers agreed through dialogue to communicate first their dissatisfaction with the task or 'instrument' and second the transformative aspect of the referents available to the group and its members. In line with Kress' "theory of the constant transformation of both resources and of subjectivity" here we see individual group members being agentive in relation to the group's resources, and in relation to the individual's own subjectivity. Don explains how teachers are forced to look seriously at the resources that are available to the individual in that transformative activity.

> Don: I think it was very much centring on the way in which the assessable elements are written... So basically, I think, all of our disagreements, to put it in its essence, were the different ways in which teachers tried to adjust for perceived weaknesses in the instrument and its criteria.
>
> I: Right.
>
> Don: Now, the last question talks about floating, sinking, Plimsoll line, weight and force. And then the assessable element out of nowhere grabs upward force. Upward force is not in the stimulus material at all. But to get an A or a B you had to use upward force, so we just ignored it. And we agreed to ignore it. But we had to dialogue the fact that we were ignoring part of the instrument because it's strop.

This data is also illustrative of the key threat to validity, of construct under-representation, as the task does not appear to adequately sample the construct of upward force that it is supposedly designed to assess.

Teacher and Student Use of Standards for Improved Learning

The standards were seen to be beneficial in that they are helpful to some teachers and their students to focus on the qualities that are assessed in the completed work and they help to address the threats to validity such as construct irrelevance or construct under-representation. This view was not shared by all teachers. Here, Carl suggests that the standards help to 'crystallize' the qualities of the work for both the teacher and the students 'to see' how the standards focus on student attention on "what they need to improve and what are their strengths" and "what we have marked". However, the use of standards for learning does not appear to be a pedagogic focus suggesting that Carl has used the discourse without fully understanding the implications for practice. This is an example of 'false clarity' in that Carl appears to see the connection between criteria and standards and learning improvement but can't see how to realise pedagogic action.

> I: Okay. So, what then … what are the benefits of using stated standards in your opinion?
>
> Carl: Um, it gives a much more lucid and crystallised idea of what this work, what the qualities of this work are and what, what perhaps the deficiencies might be as well, and it's much more based

on, um, it's much more **based on the work rather than the student**. It is also, it's also a transparent system, basically. It's a much more transparent system. And it creates much more equity between students because **it is entirely task-based** and it is the quality of the work that we are assessing at all stages and it is, and it's also very detailed. Detail is important as well. Without that detail it becomes murky and open to interpretation and, um, it **gives the students a much better idea of what they need to improve and what are their strengths**, um, and they can use those, those statements of standards and **what we have marked** and how we have used that page in order to have a much better understanding of what, further than our comments or just having a mark, of what they have achieved.

When asked about the language and the terms used to describe the standards Carl highlighted the need for specificity and began to explain the qualitative differences in the standards using his knowledge of the constructs relevant to the subject domain English.

Carl: I'd say yes, quite, quite specific. Very much task-based, very much related to the task. Um, key words, such as common curriculum elements, should be used in order to have a consistency between tasks as well, and also the specific, the specific outcomes that we are looking for the students to achieve, such as vocabulary, such as, um, and, and, and specific descriptors that show the difference between standards so that students are able to gauge in these two marks as well. So, um, for example things like, "Has explored a broad and apt vocabulary," would be something that we would see more in the A and then lower in the C we might see, for example, for a more standard result we might see, "Has used a suitable vocabulary." So quite detailed so we are able to really, in an obvious a clear way point out the quality of the work.

The key hindrances to gainful student use of standards appear to include the teacher's lack of understanding of how to integrate the use of standards and criteria within his pedagogic approaches to support student learning. In what follows priority is given to teacher talk about criteria without connection to opportunities for students to engage in application of the criteria. The pedagogy does not extend to: teacher modelling of how to use the criteria for self monitoring and improvement purposes and exemplification through illustrative samples of student work.

I: Okay. Now, so then what are your concerns about the use of stated standards?

Carl: I, my, one of my major concerns is whether or not the students can actually use them, whether they are a tool for students of whether they are a tool for, for teachers. And even with many of my attempts with my classes, for example, the students do not find much currency in them, despite the fact that we have poured through it and we have said, "This is what you need to achieve and this is exactly what an A would be. This is, this is where you would need to, this is what would allow you to achieve that standard," for example. And that still doesn't hold very much currency for the students...

Tony in contrast to Carl (see first segment below) could see more benefits for students in their use of the standards for self-assessment to identify the gains that they have made or the areas where they need to develop or focus.

Tony: Um, I can see **benefits for the kids in that they can actually see**, um, what **requirements they need to obtain a certain level**, or an A, B, C or D, so basically they can go through their work and say, "Yep, now I've done that, I've done that, I've done that, I've done that. I should be getting around about this mark when I get it back." Or they think, "Oh, geez, I should have, I could have, I haven't really done that so that's going to bring me down."

The need for consistency of the application of standards for student improvement is a recurring theme across the data sets and illustrates the acceptance by teachers of the underlying principle inherent in the policy intent of the QCAR initiative. Consider for example the clear resonances in what Carl, Tony, Ian and Cathy say regarding stated criteria and standards across these segments.

> Ian: Um, … the point of the **standards is to help, to help achieve consistency**, and I think consistency is important because **it helps students know what they need to do**. Um, the worst thing for a student, and I mean, I know from my own experience as a student a lot, the worst thing is when you have an assessment item and you don't actually know what's expected of you. You know, "Is this enough? Is this too much? How do I know?" So a specific standard that can be, that I know whoever's marking this is going to apply will help me as a student, so I can only assume it will help them.
>
> Cathy: I think it's very clear to both staff and student at the end of a task, and therefore at the end of a year and looking at all of those tasks, **they know what standard they're at**. … I've always been a big believer in standards. I just think it gives **a lot more security to the students in their learning** and to the staff in the delivery of the teaching...

It would appear from this data analysis that in the main the teachers' accounts of policy in practice align with the official policy directions and partial uptake of the QCAR initiative however there is still the need to build capacity in teachers' understanding of assessment as it connects with learning theory.

Teacher Judgements: Analytic and Holistic Approaches

The Role of Criteria and Standards

A finding from the study is that the textual representation of assessment criteria and standards – the format of what in this case is called the 'Guide to Making Judgments' – sets teachers up in particular ways to understand markers of quality. Such understanding is expected to operate at the micro level where the focus is on discrete assessable elements (criteria) linked to questions of the QCAT. It also operates at the macro level involving award of an overall grade. In judging student achievement on the centrally-devised QCATs, teachers were asked to arrive at 'on balance' judgment, this task necessarily involving attention to qualitative levels of difference.

In the QCAR initiative, priority was given initially to the micro level. To illustrate, as part of the QCAR documentation, the Guide to Making Judgments, in its original design (see Appendix 1 for Year 6 Science), and the annotated student work samples, provided meticulous specification of what teachers were to do to assess the student's work. To illustrate:

> locate the evidence in the student work for each assessable element. Match the evidence for each assessable element to a task-specific descriptor in the *Guide to making judgments*. Refer to the *Annotated student work samples* (if available) to support your understanding of the expected student response for each task-specific descriptor. (Information sheet on 'reviewing process', QSA, 2007)

Teachers were thus informed to assess by judging the component parts of the work against each element of the Guide, using annotated samples as support. Teacher judgment was in this way oriented to an analytic approach, focusing on prescribed, discrete elements. The assumption was that the process of treating each element separately, and in turn, would lead to a systematic, even regulated approach to judgment that could deliver consistency. Brief notes on obtaining an overall grade were available however no exemplars were available to illustrate the qualities expected for overall final grades of A through E. In this way the criteria were atomised at the level of the question and the standards (task-specific descriptors) that were to assist in the overall judgement for the award of a final grade remained in the background. The Guide, in its original design, adopted what we have elsewhere described as a matrix approach (Wyatt-Smith, Klenowski & Gunn, in press).

Teacher Use of Textual Referents

The recorded talk showed that teachers relied on a range of practices and referents to make the move from the parts (micro) to the whole (macro) in arriving at a judgment. These included giving priority to the annotated samples, referring only to the Guide as a secondary source of information; giving each of the criteria a numeric sub-score and then totalling the sub-scores to arrive at an overall grade; parcelling out the marking to different teachers to judge certain sections of the paper only and then passing the responsibility for overall judgment to another party (usually a senior teacher or curriculum leader) to combine the judgements on the separate criteria into a composite grade. As an example of the teacher use of samples Martin can be heard indicating his reliance on the samples as a way to cue himself into 'categories' primarily to inform judgment. Rhonda is also reliant on samples to inform her judgements.

> I: So what process did the teachers go through in terms of reaching their judgments?
>
> Martin: In terms of reaching their judgments the, um, the teachers had pretty carefully read through all of the information provided by, um, QCAR and they had made sure that they, um, firstly looked at the, they had a look at the tests themselves, they looked at the sample responses and they tried to, ah, align their judgments with the sample responses.
>
> I: And then sort of mark those on the back with all of the, with what they call the Guide on the back?
>
> Martin: Yes, which we found to be both helpful and unhelpful. In some cases, see, answers were, ah, responses were very explicit and they **fitted into a category** very easily. In other cases, students answered in different ways which didn't make **categorisation** easy.

Rhonda also explains how in assessing the mathematics task problems emerged for her when combining components of the task **to award an overall grade**. It was in this context that the teacher referred almost entirely to the sample responses to see how the grades were awarded. However, she found these samples did not account for all possible variations so that: "It was difficult to give a C when one question was fully right and one was fully wrong".

> I: Yeah. So, how did you use the materials and, and what was the way you went through the task?
>
> Rhonda: I basically used the **task-specific descriptors** [standards]and the **question numbers** relevant to it and **then graded it**, as I told you before, I have done a lot of senior Maths marking where we mark on criteria, so that helped me a lot to do it, so I feel that was quite straight-forward and easy that way. Though, when marking individual questions, it was a bit difficult to give them a C when one question was fully right and one question was fully wrong and those types of things and we are to give how much value? That was a problem there.

Both Martin and Rhonda touch on the complexity of judging, suggesting their interest in materials that could make 'categorisation easy'. There is some suggestion that the matrix approach of the guide hindered the award of an overall grade because of the extent to which the QCAT and the assessable elements atomised the teachers' approach to judgement. A more experienced Head of Curriculum spoke of how he drew on his evaluative experience – another way to see - in working with teachers in his department to show how strengths and limitations could be evidenced in a piece of work and how these could facilitate on balance judgements.

> Ben: I think what, ah, when we had a look at this QCAR products, when we had a look at those and we saw the student responses to that, it was interesting in that with the A description, the B, I think that was a very good method of showing what is required. … But I had a chat with one of my other colleagues before and he was saying, "You know, this, this question here, that's an A standard," and then he turned the page and showed a B standard and the wording was virtually the same. I had to point out to him that what the standard was showing you wasn't just for that question but for the whole paper. So, you know, the B standard, to my way of thinking, for both questions, the two answers for both questions were a very good answer, but it was in the other

questions where the B standard would have come out, not necessarily just in that one question. I think that's something that needs to be pointed out[ii]...

Forward Research Directions

The research findings of this study have been used to inform the next stage of the QCAR implementation. For example, an alternative design for the Guide is being trialed in 2008 (See Appendix 2), with the alignment of standards to criteria shown graphically in an approach applied previously in the work of New Basics (Klenowski, 2007). This is a focus of our continuing research with the QSA.

A future direction for the research presented here is to focus on how to develop teacher assessment capacity to use criteria and standards to inform judgement for teaching, learning and reporting purposes. This will entail a professional development strategy that includes assessment principles, models of judgement and accompanying resources. A focus would be on teacher use of standards at both task and discipline levels and would be applicable to both National and State curriculum and assessment priorities. The strategy could draw on the work of Smith (1989) which involved trialing the theorising of standards developed by Sadler (1987) in the case of Senior English; the body of writing on ways to make teacher assessment dependable (Harlen, 2004; Klenowski, 2008; Sadler, 2008), and related empirical research on judgment practices (Wyatt-Smith and Castleton, 2005; Cooksey, Freebody and Wyatt-Smith, 2007; Klenowski and Adie, in press). Three components of this strategy include:

1) elaborated guidelines about on-balance judgment processes, focusing attention on how teachers consider the characteristics of the work against each of the specified properties of the standards (e.g. A-E) and analyse the configuration of these properties to determine those that are dominant in the student work.
2) exemplar student work (on a task, extended to a portfolio) indicative of the standards, that is to illustrate 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 C. The role of these materials is to illustrate different ways of satisfying the stated requirements of the criteria and the standards. In effect, they serve to convey to teachers that it is reasonable to expect student work to show different performance profiles in relation to any set of given criteria. Also to the fore is the message that teacher judgment using standards written as qualitative descriptors is not technicist in nature and that the application of such standards requires recognition of trade-offs or compensatory factors.
3) descriptive reports of student achievement accompanying the exemplars to given insight into the factors which influenced the overall judgment and the final achievement award. Such reports provide information about the decision-making processes that the teacher relied on to arrive at an overall judgment, including specifics about how trading-off of perceived strengths and limitations was treated.

This strategy carries forward the understanding that standards written in qualitative terms, such as those presented in the QCAR materials (see Appendices 1 and 2) represent mental constructs and 'can have their interpretation circumscribed, more or less adequately, only by usage-in-context' (Sadler, 1987, p.206). Further, it concentrates attention on how 'specifying and promulgating Levels of Achievement as standards must address practical considerations' (p.196). In the context of QCAR, this relates back to the usefulness of the Guide to Making Judgments and the discipline standards to teachers in their attempt to identify standards.

The strategy has the potential to connect teachers' ways of working with criteria at levels to a necessary focus on the match between the work (the evidence) and the standards against which it is to be assessed. Drawing on the findings of Smith's (1989) study, it would be useful also to include a front-end statement or preface emphasising two points. First is the distinction between achievement and non-achievement and attitudinal consideration, including diligence and disposition, and that the determination of a grade must depend on a decision of its match to the stated standard. Second would be information about the intended functions of the exemplar materials, be it in the form of stand–alone completed tasks and related student work samples, or portfolios including a range assessment tasks, emphasising their

illustrative (rather than prescriptive) nature, highlighting that other ways of meeting the standards are also possible.

The findings of the study to date have shown that a common interpretation of the standards, at the level of chosen discipline tasks, is in development. Also clear is that in the main, teachers have not connected the Guide to Making Judgments to the standards developed as part of the QCAR initiative for judging achievement at the discipline level. That is, there is no conceptual bridge linking the Guide to discipline standards. Further, it is worth emphasising that while it is widely recognised that discussion among teachers regarding the evidence depicting the qualities of standards is fundamental, our observation is that such discussion will not necessarily occur in the absence of policy direction. We suggest that this observation holds, even if individual schools are proactive and make time provision for moderation linked to professional learning.

References

Cooksey, R.., Freebody, P. and Wyatt-Smith, C.M. (2007) Assessment as Judgment-in-Context: Analysing How Teachers Evaluate Students' Writing. *Educational Research and Evaluation*, *13*(5), 401-434.

Crooks, T. J., Kane, M. T. and Cohen, A. (1996) 'Threats to the valid use of assessments,' *Assessment in Education: Principles, Policy and Practice,* 3(5), 265-285.

Frederiksen, J. R. and White, B. Y. (2004) Designing assessments for instruction and accountability: An application of validity theory to assessing scientific inquiry, in: M. Wilson (Ed) *Towards Coherence Between Classroom Assessment and Accountability,* The 103rd Yearbook of the National Society for the Study of Education Part 2, Chicago: National Society for the Study of Education.

Harlen, W. (2004). *Can assessment by teachers be a dependable option for summative purposes?* Paper presented at General Teaching Council for England Conference, 29 November, 2004: London.

Harlen, W. (2005) Teachers' summative practices and assessment for learning – tensions and synergies, *The Curriculum Journal,* 16(2), 207 - 23.

Klenowski, V. and Adie, L. (in press) 'Moderation as Judgement Practice: Reconciling System Level Accountability and Local Level Practice', *Curriculum Perspectives.*

Klenowski, V. (2007) Evaluation of the effectiveness of the consensus-based standards validation process. Townsville: DETA. Available online: http://education.qld.gov.au/corporate/newbasics/html/lce_eval.html

Klenowski, V. (2008) 'A Call to Honour: Teacher Professionalism in the Context of Standards Referenced Assessment Reform', in A. Luke and K. Weir *Development of a Set of Principles to Guide a P-12 Syllabus Framework: A report to the Queensland Studies Authority,* Brisbane: Queensland Studies Authority.

Kress, G. (2000) "You've Just Got to Learn How to See": Curriculum Subjects, Young People and Schooled Engagement with the World, *Linguistics and Education,* 11, (4), 401-415.

Messick, S. (1989) 'Validity' in R. Linn (ed.) *Educational Measurement (3rd edn),* New York, NY: American Council on Education and Macmillan, 13-103.

National Curriculum Board, (2008) National Curriculum Development Paper, Accessed www.ncb.org.au

Nuttall, D. (1987) 'The Validity of Assessments,' *European Journal of Psychology of Education',* 11, 109-18.

Queensland Studies Authority (2008) P-12 Assessment Policy, Brisbane: Queensland Studies Authority.

Sadler, D. R. (1987) Specifying and promulgating achievement standards, *Oxford Review of Education,* 13(2), 191-209.

Saunders, L. (2005) 'Policy Research', Presentation at the Institute of Education, University of London.

Shepard, L. (2000). The Role of Assessment in a Learning Culture. *Educational Researcher,* 20(7), 4-14.

Smith, C. (1989) A study of standards specifications in English. Master of Education (Unpublished). Brisbane: University of Queensland.

Wilson, M. (2004) (Ed) *Towards Coherence Between Classroom Assessment and Accountability,* The 103rd Yearbook of the National Society for the Study of Education Part 2, Chicago: National Society for the Study of Education.

Wyatt-Smith, C., & Castleton, G. (2005). Examining how teachers judge student writing: An Australian case study. *Journal of Curriculum Studies,* 37(2), 131-154.

Wyatt-Smith, C., Klenowski, V. and Gunn, S. (in press) The centrality of teachers' judgement practice in assessment: a study of standards in moderation, *Assessment in Education: Principles, Policy and Practice.*

[i] The project is funded by the Australian Research Council in collaboration with Industry Partners, the Queensland Studies Authority and National Council for Curriculum and Assessment (The Republic of Ireland).

[ii] See Guide to making judgments – Year 6 Science Appendix 1