Abstract and Paper for AARE 2008 Conference
Paper WYA081029

Title: Standards-Driven Reform Years 1-10: Moderation an Optional Extra?[1]

Authors: Val Klenowski
Queensland University of Technology
val.klenowski@qut.edu.au

Claire Wyatt-Smith[2]
Griffith University
claire-wyatt-smith@griffith.edu.au

Key words: assessment, standards-based reform, judgement, moderation

Abstract

While externally moderated standards-based assessment has been practised in Queensland senior schooling for more than three decades, there has been no such practice in the middle years. With the introduction of standards at state and national levels in these years, teacher judgement as developed in moderation practices is now vital. This paper argues, that in this context of assessment reform, standards intended to inform teacher judgement and to build assessment capacity are necessary but not sufficient for maintaining teacher and public confidence in schooling. We will argue that teacher judgement is intrinsic to moderation, and to professional practice, and can no longer remain private. Moderation too is intrinsic to efforts by the profession to realise judgements that are defensible, dependable and open to scrutiny. Moderation can no longer be considered an optional extra and requires system-level support especially if, as intended, the standards are linked to system-wide efforts to improve student learning. In presenting this argument we will draw on a current ARC funded study with key industry partners (the Queensland Studies Authority and the National Council for Curriculum and Assessment of the Republic of Ireland). The data analysed will include interview data and some teacher talk during moderation. The analysis will highlight the issues that emerge in moderation meetings that are designed to reach consistent, reliable judgements. Of

[1] Acknowledgment

The authors wish to acknowledge that this is an ARC funded Linkage Project and involves support provided by the Queensland Studies Authority (QSA) as our Industry Partner and the National Council for Curriculum and Instruction of the Republic of Ireland. We also wish to acknowledge the significant contributions from our researchers Lenore Adie and Stephanie Gunn.

[2] While names are expressed alphabetically both authors have contributed equally to the writing of this paper.

interest are the different ways in which teachers talk through and interact with one another to reach agreement about the quality of student work in the application of standards. There is also some emerging evidence of differences in the way that teachers make compensations and trade-offs in their award of grades dependent on the subject domain in which they teach. The paper will conclude with some empirically derived insights into moderation practices as 'policy' and 'social' events.

Standards-Driven Reform Years 1-10: Moderation an Optional Extra?

Introduction

In October 2008 the Australian Curriculum Assessment and Reporting Authority (ACARA) Bill was presented to parliament to establish a single authority responsible for the functions of curriculum, assessment and reporting at the national level. This authority will work with the National Curriculum Board, established in January 2008 by the Federal government, to set the core content and achievement standards in Mathematics, Science, History and English from Pre-school to Year 12. The National Curriculum is to be developed by 2010 and in 2011. A second phase will extend the development to include Languages and Geography. 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.

Most recently the National Curriculum Board has issued a proposal for discussion regarding the use of 'achievement standards' that are intended to indicate the quality of achievement that is expected and to provide the basis for judgements about the quality of students' work (National Curriculum Board, 2008). The purposes that the 'achievement standards' must fulfil are outlined as follows:

- Make clear what quality of learning (knowledge, understanding and skills) is expected to be achieved;
- Provide helpful language with which teachers can discuss with students and their parents the students' current achievement level, progress to date and what should come next and
- Help identify students whose rate of progress puts them at risk of being unable to reach satisfactory achievement levels in later years (ibid).

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 policy, particular assessment initiatives and practices. The research evidence will have relevance for the implementation of standards-based assessment not just locally but will extend to include national and international efforts. It is for this reason that our industry partners include the National Council for Curriculum and Assessment in the Republic of Ireland.

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 stated 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, assessment bank) rather than processes. This study 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).

Theoretical Underpinnings of 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 by more than one researcher. The theoretical lens used to read and analyse the data has drawn substantially on the work of Sadler (1987). Standards-referenced assessment relies on teacher judgement that can be made dependable if standards are 'promulgated' in appropriate forms and teachers have the conceptual tools and professional training. The methods for promulgation include: numerical cut-offs, tacit knowledge, exemplars and verbal descriptors.

Sadler argues that when multiple criteria are used in assessment it is more viable to work with 'fuzzy composite standards' (Sadler, 1987) in contrast to sharp standards characterised by precise boundaries and numerical cut-offs. "[V]erbal descriptions are always to some degree vague or fuzzy" (Sadler, 1987: 202). The fuzziness of grade boundaries comes from the interpretation on a particular criterion underlying the standards scale, and the interpretation of the different levels separating one standard from another. Both these interpretations depend on context. A competent assessor using verbal descriptors to judge student work will make "compensations

and trade-offs … to allow for intercorrelations among the criteria and for the multiplier effect of some criteria on others" (Sadler, 1987:206). The central task for the assessor in grading student work then is to find the 'best fit' with the grade description knowing that no description will be a perfect match.

Sadler makes reference to Wittgenstein (1967/1974) who understood that precise definitions for a lot of the everyday concepts was impossible. In his view they comprised "a complicated network of similarities, overlapping and crisscrossing: sometimes overall similarities, sometimes similarities of detail … 'family resemblances.'" A combination of language and experiences with real phenomena help 'sketch in' meanings. Sadler (1987: 206, our emphasis) concludes, "the verbal description of a standard (the standard itself being an abstract mental construct) can have its interpretation, circumscribed, more or less adequately, *only by usage in context.* The concrete existential referents that make up the context are essential to its proper interpretation."

A sociocultural view of learning sees learning as socially negotiated and embedded within a cultural community (Murphy and McCormick, 2008). This theoretical lens has been added to our reading of the data, as we understand that learning is both a process of 'becoming competent and belonging' (ibid: x). In this study we are analysing how teachers become competent assessors in their consistent use of standards. From this view 'identity and knowledge are seen as interdependent'. The importance of context is apparent from this sociocultural analysis of the interview and conversation data. The moderation meetings focus on the quality and standard of the student work under discussion. As these teachers participate in such contexts they draw on a range of concrete referents and social and tacit knowledge. At the same time they are learning in practice (about the standards) and negotiating an identity (as an assessor) (ibid).

Reflecting the socio-cultural framing of the project, are understandings about language itself as inherently social and cultural (Kress, 2000). This approach is consistent with the understanding that the terms in which standards are written come to acquire meaning over time, through use in contexts, including institutional and domain contexts and interaction contexts such as moderation. The latter can take place incidentally or more formally in a structured manner through system and school initiatives.

In this paper we draw on the notion that the teacher or sign-maker is "constantly transformative of the set of resources of the group and of her/himself" (Kress, 2000: 401). '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 the data reveals that there is interaction with several modes of representation and communication and as such is multimodal. In the analysis of the interactions we identify a specialisation of representational and communicational modes.

To date the study has identified an empirically derived framework of categories of referents that teachers have used to confirm or challenge proposed grades during the moderation meetings (Wyatt-Smith et. al. in press). 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. Data analysis to date shows that 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 there are six requirements for social moderation to work effectively. 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 is that there are "teacher discussions of the quality of the assessment instruments and the standards of student work" (QSA, 2008: 1). 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, trialed and developed by the QCAR team using classroom teacher input and feedback. The tasks have been written to provide 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 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. If teachers vary in their interpretation of the construct of interest that the task is designed to assess then there will be variation in the judgements which will impact on the reliability of the grades awarded. Another example of construct irrelevant variance would be when the teacher's prior knowledge about a student influences the judgement made regarding the grade. To illustrate, if a teacher views a student in terms of their identity as "an A student", and draws upon this knowledge when awarding the grade for the QCAT then this is irrelevant to the construct that is being interpreted in the QCAT. That is, the focus should be on the work that the student has completed for the particular QCAT not what has been accomplished previously.

Discussion and Findings

What follows is an analysis of data that illustrates how teachers are 'learning to see', or becoming competent in their use of standards, from their interpretation of the grades 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 an oral mode of commentary, critique and ratification, as well as written comments recorded on student samples, annotations entered on the Guide, usually involving 'shorthand markers' of quality such as ticks or highlighting, and in some cases a written comment to accompany the letter grade.

Standards Informing Judgement

Standards have been categorised into five types (Maxwell, 2008) but for the purposes of this paper the two that are most relevant are those concerned with differentiated levels of performance. Standards described as 'arbiters of quality', indicate relative success or merit, the other type is described as 'milestones' and are indicative of progressive or developmental targets. These two types differ in focus and timeframe. The former are merit standards that can apply to a single assessment event, such as a completed task or the QCAT, while the latter are developmental standards that inform judgements that can be made along a continuum of learning over time, such as for a portfolio of work (Maxwell, 2008).

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. This is a merit standard or an 'arbiter of quality' so the focus in the judgement is on relative success or merit of the student's response. The importance of context, the interactions and the transformative nature of the available resources become apparent:

> Chris: I think I'll be interested to see, … the most important part for me … 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 grading of the student's work and his expectations will either be confirmed or challenged. 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 due to the ambiguity associated with the construct to be assessed. 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" (op cit) 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. The important finding here is the need for greater clarity and explication of construct definition in the task for teachers.

Use of Standards for Improved Learning

The standards were seen to be beneficial in that they help 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. All teachers did not share this view. 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 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 part of Carl's pedagogy suggesting that he 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 does not seem to know how to realise this pedagogically.

> I: … what are the benefits of using stated standards in your opinion?
>
> Carl: Um, it gives a much more lucid and crystallised idea of … what the qualities of this work are and what, … the deficiencies might be as well, and … it's much more based on the work rather than the student. … it's also a transparent system … 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 … Detail is important as well. Without that detail it becomes murky and open to interpretation … 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 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.

Carl does not appear to know how to assist students in the use of standards for learning by incorporating assessment for learning strategies. His talk about criteria does not include the provision of opportunities for students to apply the criteria. His 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: … one of my major concerns is whether or not the students can actually use them, whether they are a tool for students or 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 pored through it and we have said, "This is what you need to achieve and this is exactly what an A would be. … this is where you would need to,

> this is what would allow you to achieve that standard," … And that still doesn't hold very much currency for the students...

Tony in contrast to Carl can 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. … 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?"

> 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

The textual representation of assessment criteria and standards or the Guide ('Guide to Making Judgements') directs teachers 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 that involves the award of an overall grade. In judging student achievement on the centrally-devised QCATs, teachers were asked

to arrive at 'on balance' judgement, this required attention to qualitative levels of difference.

In the QCAR initiative, priority was given initially to the micro level. The Guide, in its original design (see Appendix 1 for Year 6 Science) adopted a matrix approach, 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 Judgements. 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, represented in matrix format, using annotated samples as support. Teacher judgement 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 judgement 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 (A – 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.

Discipline differences and their role in shaping judgment

The talk recorded in interviews and moderation meetings has also brought to light the contributions of discipline knowledge in teachers' expectations about how judgment should properly occur. Specifically, the data show how teachers of English, for example, enact judgment processes in significantly different ways from teachers of mathematics and science, even when they are using a Guide to making judgement that has a common design across the three discipline areas. This points to how teachers' ways of recognising achievement can be traced directly to particular constructs of knowledge.

Consider, for example, how in answer to the question about the procedures and processes relied on to achieve consistency of teacher judgement in science, Morris refers to the benefits of teachers 'following a similar marking key'… a very detailed marking key'. For teachers of mathematics and science, the recurring interest was in how tightly specified marking information was essential in regulating judgment, and moreover, that questions should be accompanied by marks that indicated their relative importance in the assessment item. In the following interaction, a teacher of primary maths and science was asked to elaborate on his statement that 'the key thing [for achieving consistency of teacher judgement] was to get a common understanding of the standard', to which he said:

> *In the, in the ideal world of education we have the task, we have its criteria and we have a sample response. Ah, in maths and science exams that's easy, you have a marking sample. And the other teachers can look and go, "Okay, so there's two marks for that question and that's broken down by identifying the formula". So it's not just seeing the marks. I would also expect teachers … to also be able to see the part marks"*

In this segment the teacher can be heard associating criteria, a sample response and a marking schedule that stated how marks, including part marks, were to be awarded. The concern is with the component parts of the work to be assessed, though missing from the talk is how marks and standards connect in practice or conceptually.

Overall, the prevailing assumption in the recorded talk of maths and science teachers was that marks were used not only to regulate how judgment should occur, but that marks were the building blocks for arriving at overall judgment. In short, the teachers' assessment gaze was with the parts and that the whole necessarily represented the sum of the parts. Moreover, there was the expectation that judgment practice could be wholly prespecified, some teachers talking about the merit of prescribing the answers to be recognised for half and quarter marks. One teacher spoke of this saying how teachers 'really have to be more consistent in, where some teachers might be inclined to give a half-mark, um, where others might only give a quarter mark. That kind of thing. I usually find that that's the area that needs the most attention.'

Further, from maths and science teachers there was a level of reported discomfort in using standards that they regarded to be 'open to interpretation'. In the words of one teacher, the matrix approach to specifying standards, referred to above, ' is obviously open to interpretation'. Continuing on, the teacher characterised the matrix as 'probably not as objective as it needs to be', commenting favourably on how in the school, 'we make up a proforma – a matrix for marking and marking criteria'. It was as though standards written as verbal descriptors, representing fuzzy standards, needed to be fixed and that a marking criteria using numeric scores needed to be developed to stabilise the meaning of the terms used in the standards.

Such practices, while common in science and mathematics, were not evident in English. One teacher captured the widely reported judgment stance in English as follows:

> *We're always looking for a global assessment. We're not looking at a precise number in English, so you're always looking at a global… .*

Overall, English teachers tended to move from the whole to the part, preferring to regard the work to be assessed in its entirety, before fixing on particular aspects of performance. Further, collectively they voiced concern about 'a danger of being too detailed' in specifying criteria and standards, as indicated in the segment below:

> *They need to be, well, for English teachers who interpret every single letter, let alone every single word, they need to be explicit and they need to be up-*

> *front and they need to be well-understood by people prior, perhaps, to doing anything with them. Um, but you never really understand what they're about until you are grading or you are using them. So, until you see them in operation it's hard to know, but there is a danger of being too detailed and almost verbose with what you're trying to do…. The standards have to reflect really, it hones in therefore on what it is you're really assessing.*

In this segment we hear the teacher disclosing how knowledge of the standards is acquired through use, and warning against an overly detailed approach to how they are formulated. There is also the telling statement that the standards 'have to reflect' what is 'really' being assessed.

Such an observation opens up for consideration how maths and science teachers, and English teachers, have different expectations of how standards function in informing judgment. For the former group, the standards, in conjunction with numeric scores, are expected to regulate teacher attention, while for English teachers, given standards acquire meaning through use. It is as though teachers try on the 'fit' of the standards for student work as an evaluative experience, with the terms in which the standards are written acquiring meaning within a marking occasion, and from one occasion to the next.

The final observation in this section of the paper concerns the longstanding distinction between objectivity and subjectivity and the prevalence of these terms in teacher talk about disciplines and judgment practices. The study has provided some considerable evidence that maths and science teachers are more likely to claim that objectivity of judgment is realised through numeric scoring. English teachers report valuing holistic judgment, taking this to be subjective judgment practice. This is clearly voiced in the segment below:

> *T: English is inevitably going to have some subjectivity in it, um, and then I think this is the challenge with defining our criteria in each of the standards… as I said, we want to be, we want to be as objective as possible, and um, that's very difficult in English. It's just not a you know//*
>
> *I: It's not an objective subject.*
>
> *T: That's right. You know, it's not a quantitative, you know, assessment. It's qualitative and quantitative, you know. It's the holistic judgement of the piece and, you know, you've got to find a balance.*

The reference to English teachers wanting to be as objective as possible can be heard to signalling a commitment to fairness in grading. It can also be heard as signalling that while English teachers aspire to objectivity, in the opinion of this teacher, this is not realised through adopting what she refers to as wholly regulated, quantitative approach.

The discussion to this point suggests that the ecology of teacher judgment is shaped as much by assumptions about the nature of knowledge as it is by any given or

prescribed approaches to judgment using a Guide or set criteria and standards. The study has also shown how judgment practice can be shaped by textual materials teachers are given, as discussed next.

Teacher Use of Textual Referents

As suggested above, 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 judgement. 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 judgement to another party (usually a senior teacher or curriculum leader) to combine the judgements on the separate criteria into a composite grade. Martin can be heard indicating his reliance on the samples as a way to cue into 'categories' primarily to inform judgement. Rhonda is also reliant on samples to inform her judgements.

> I: So what process did the teachers go through in terms of reaching their judgements?
>
> Martin: In terms of reaching their judgements 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 judgements 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[3] and the question numbers relevant to it and then graded it, as I told you before, I have done a

[3] The teacher has used the incorrect term in this context. The assessable elements identified the constructs and their relationship to the questions of the QCAT not the task-specific descriptors which were in fact the standards. That is the task-specific descriptors are the standards and the assessable elements are the constructs and/or criteria.

> 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 (HoC) 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]...

This HoC has demonstrated his awareness that in making an on balance judgement the approach requires 'best fit' rather than 'perfect match' in terms of the teacher's interpretation of the evidence that demonstrates the student's understanding of the construct being assessed at a particular standard, in this case a 'B'.

Forward Research Directions

These research findings have informed the next stage of the QCAR implementation in that an alternative design for the Guide is currently being trialed (See Appendix 2). The alignment of standards to criteria is shown graphically on continua as in the work of New Basics (Klenowski, 2007). This forms part of our continuing research.

This study also aims to report on how to develop teacher assessment capacity in the use of criteria and standards to inform judgement for teaching, learning and reporting purposes. A professional development strategy that includes assessment and moderation principles, judgement approaches and accompanying resources will be developed. Teacher use of standards at both task and discipline levels for

application at National and State curriculum and assessment priorities will be the focus. This is particularly significant given the recent proposal for discussion from the National Curriculum Board of supplying teachers with annotated samples of student work that set out the basis for the assessment as a helpful approach to assist teachers in the consistent use of verbal descriptors in the award of grades both within and across schools.

The strategy that is emerging from this research draws on the work of Smith (1989) which involved trialing Sadler's (1987) theory of standards 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 judgement practices (Wyatt-Smith and Castleton, 2005; Cooksey, Freebody and Wyatt-Smith, 2007; Klenowski and Adie, in press).

First, elaborated guidelines are needed about on-balance judgement 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 analysis of the configuration of these properties to determine those that are dominant in the student work. As is emerging in our research the construct definition of the task needs to be explicated to assist teachers in their judgement approach in terms of trade-offs and compensations to reach an on balance or holistic judgement.

Second, exemplar student work (on a task, extended to a portfolio) indicative of the standards is needed to illustrate a particular achievement level (A-E). While these could be exemplified as within-band level, they could be more usefully chosen to illustrate the absolute minimum requirements for work judged to be at 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 judgement 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. That is, there is no 'perfect match' with the verbal description rather a 'best fit' approach is needed in making a judgement in the award of a grade.

Third, descriptive reports of student achievement accompanying the exemplars will give insight into the factors that influenced the overall judgement and the final achievement award. Such reports provide information about the teacher's decision-making in reaching an overall judgement, including specifics about the trading-off process of perceived strengths and limitations. They also highlight the construct properties that help determine the direction for the trade off in their judgement as these constructs represent the desirable features that cannot be undervalued.

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: 206). Further, it concentrates attention on how 'specifying and promulgating Levels of Achievement

as standards must address practical considerations' (ibid: 196). In the context of QCAR, this relates to the usefulness of the Guide 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. The construct that the task is designed to assess needs to be defined at the outset so that teachers in their interpretation can more reliably focus on appropriate evidence in the student work rather than succumb to their tacit knowledge of the student. At the outset of the instructions to assessors it is worth emphasising the distinction between achievement and non-achievement and attitudinal consideration, including diligence and disposition. That is, the determination of a grade must depend on a decision of its match to the stated standard. Second, the intended functions of the exemplar materials, as stand–alone completed tasks and related student work samples, or portfolios including a range of assessment tasks, should emphasise their illustrative (rather than prescriptive) nature, highlighting that other ways of meeting the standards are also possible (Smith, 1989).

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 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 Judgement-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. (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.

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

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.

Maxwell, G. S. (2008) *Setting standards: Fitting form to function*, Paper presented at the 34th IAEA Annual Conference, Cambridge UK.

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

Murphy, P. and McCormick, R. (2008) Knowledge and Practice: Representations and Identities, London: Sage.

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 judgements – Year 6 Science Appendix 1