External peer review of assessment: an effective approach to verifying standards?

Abstract

There is growing international concern to regulate and assure standards in higher education. External peer review of assessment, often called external examining, is a well-established approach to assuring standards. Australian higher education is one of several systems without a history of external examining for undergraduate programmes that is currently considering the approach. What can entrants to external examining at that level learn from the UK higher education system's long history of external examining? To that end, this paper reports on a mixed methods research project designed to investigate current practices in how academic standards are conceived, constructed, and applied by external examiners and debates the implications of the findings for the development of external examining in other countries. The findings suggest that the potential of experienced peers in a subject discipline to provide the assurance of standards is limited. It concludes by presenting various possible enhancements that might be considered.

Keywords: external examiners; quality assurance; standards

Sue Bloxham[a1]**, Jane Hudson**[b]**, Birgit den Outer**[b]**, Margaret Price**[b]

[a] Faculty of Education, Arts & Business, University of Cumbria, Lancaster, UK.

[b] Department of Business and Management Oxford Brookes University, Oxford, UK

[1] S.Bloxham@cumbria.ac.uk

External peer review of assessment: an effective approach to verifying standards?

Academic standards are fundamentally reference points for what students should know or be able do (Price, 2005; Sadler, 2007; Bloxham, Boyd & Orr, 2011). Processes of *quality improvement* are designed to maintain or enhance such standards and *quality assurance* is designed to demonstrate their existence. An emphasis on assuring standards in higher education is not new (Krause et al. 2013), but it is increasingly becoming a global phenomenon (Barrie, Hughes, Crisp & Bennison, 2014) in both well-established and newer university systems. International competition has placed pressure on universities to be more accountable (Dill & Beerkens, 2012) and improve the protection for interested parties in higher education such as students and employers.

A well-established approach to quality assurance of standards in a number of jurisdictions is a system of external peer reviewers of assessment, generally referred to as external examiners. In a context where several other countries, groups of universities and disciplines are contemplating introducing such an approach, it is useful to establish the potential effectiveness of existing external examiner methods. To that end, this article reports on a research project designed to investigate current practices in how academic standards are conceived, constructed, and applied in external examining processes in the UK. It will discuss the implications of the findings for the use and design of external examining more widely.

Background

There has been a recent drive to assure the standards of graduate outcomes and the achievement of comparability across universities and countries (Barrie et al., 2014 p. 19). These developments are reflected in a range of policies and projects in the USA, the AHELO

study and the cross national Tuning project involving Europe, South and North America, Africa, and Russia (Krause et al, 2013). For the most part, these projects focus on defining (or tuning) standards through their explicit articulation, for example by aligning qualification frameworks and disciplinary standards. At the national level, external oversight of standards is taking place through national frameworks for describing and safeguarding award standards and regulating quality systems, for example the UK *Quality Code for Higher Education* and the Australian *Qualifications Framework (AQF).* These frameworks include elements such as statements of graduate outcomes, qualification descriptors and benchmarks for disciplinary or professional standards. Such assurance of standards is also apparent in the use of programme learning outcomes, scrutiny of assessment methods, the use of criteria and rubrics and the internal and external moderation of assessment.

However, these expressions of quality assurance can be conceptualised as 'process' standards, part of quality management, whereas 'academic standards' are defined, for example by UK and Australian quality bodies, as 'output' focused; that is levels of achievement that must be reached to obtain an award. Academic standards are therefore only really demonstrated through academic attainment as revealed through performance in assessments (Bloxham & Boyd, 2012). Consequently, one tool used for assuring academic standards within and across universities in several countries is a system of external examiners because they focus their efforts on student outputs through scrutiny of assessment design and completed student work. Such an approach, or adaptations of this method, are also under consideration by other university systems. One example is in Australia. The December 2008 *Review of Australian Higher Education* (Bradley report) marked a renewed focus on enhancement and accountability regarding the graduate outcomes of the growing student population following concerns regarding the assurance and comparability of standards in relation to external reference points (Barrie, et al. 2014). It advised that more explicit

indicators should be developed to directly assess and compare learning outcomes (Bradley, Noonan, Nugent & Scales, 2008). Since that time, threshold learning outcomes for many subject disciplines have been developed, similar to the UK's benchmarking statements, which set out the minimum standards for graduation in a discipline. A 2013 consultation document (Australian Government, 2013) also advocated a process for assuring standards combined with periodical external peer review of assessment. That consultation has yet to be turned into policy (April 2014) but, in the meantime, several contrasting models for inter-institutional peer review of assessment have been explored including the Group of Eight Quality Verification system, the Achievement Matters external peer review of accounting learning standards (Watty et al. 2013) and a large 'proof of concept' project using an inter-institutional blind peer review of assessment methodology (Krause et al, 2013). An Australian *Peer Review Network* has been established[2]. These external examining and moderation projects include novel 'calibration' methods concerned with professional learning as well as more traditional, UK-type, external examiner processes (Deane & Krause, 2012). Within these developments, there is an assumption that variation between reviewers can be tempered by the provision of common external reference points such as disciplinary threshold learning outcomes, such that they 'boost… the objectivity or trustworthiness of external reviewer judgements' (Barrie et al. 2014, p. 24). This is commonly agreed to be important although there is also a recognition that external reference points alone have limited power to ensure comparable judgement without other community processes to calibrate individuals' judgement. Overall, the issue of whether and how to make use of external examining in Australian universities is very much a live issue with debate continuing regarding the balance between a light touch and more extensive professional learning approaches (Barrie et al. 2014). Any university or university system considering an extension of external examining,

[2] (www.utas.edu.au/serru/nprn).

whatever the method, would be wise to draw on studies of its effectiveness in jurisdictions where it is well established. This paper provides such a study.

External examiners and standards

External examiners are routinely used in many countries for the assessment of post-graduate work. However, there are university systems, such as the UK, where external examining is in widespread, often mandatory, use as a key tool in assuring assessment standards in undergraduate education (see QAA 2011 for detailed expectations of the process). A key aspect of this external examining is that it draws on disciplinary expertise within the discipline in order to reflect the epistemological differences in assessment practices across subject areas (Trowler, 2009, Barrie et al. 2014). Although roles vary, a primary task is the scrutiny of both assessment/ examination tasks and examples of student performances. Studies of the effectiveness of external examining in the UK have identified a number of operational criticisms over the years (Bloxham & Price 2013) but have not challenged the assumption that the basic concept of inter-institutional peer review is effective. Investigations into the role have not focused on the capacity of examiners to hold and consistently apply a shared knowledge of academic standards or tested the existence of effective processes to support the development of consensus in standards (Bloxham & Price 2013).

This omission is important because research on academic judgement and grading from a range of disciplinary and epistemological perspectives broadly shares negative findings regarding the consistency of academic standards in higher education assessment (O'Hagan & Wigglesworth, 2014). Researchers are increasingly drawing on a socio-cultural framework to investigate and explain academic judgement. In sociocultural theory, professional learning is

not something that is acquired passively from instructors and mentors, but is perceived as something that is jointly created between the professional and their social environment. Learning is mediated by the artefacts and language of social action and is a process of enculturation in which new knowledge may be created (Wenger, 1998). Consequently, the socio-cultural approach conceives of academic judgement in assessment as a socially situated interpretive act where the meaning of standards is constituted through the shared practice and dialogue which takes place in the social, cultural and political contexts concerned (Shay, 2004). In this way, 'calibration' (or shared understanding) of standards is a social rather than a technical process. Such research consistently emphasises the individualised, tacit, interpretive nature of standards. Assessors' judgements are influenced by their experience, values, habits of mind, norms of student work and knowledge of students. They focus on different aspects of student work and they make limited use of codified standards which, in themselves, pose problems of shared interpretation. Overall, assessors' inconsistency and unreliability is well documented (see summary of research and references in Bloxham & Price 2013). Studies in the fields of psychology and cognition also demonstrate the lack of consistency in academic judgement caused by a number of characteristics of complex decision-making (Brooks, 2012).

Despite these findings, academics receive relatively little induction or training in relation to assessment and standards, learning for the most part through personal experience (Yorke 2009). Efforts to achieve consistency through determination of explicit statements of standards have been shown to have shortcomings. Krause et al. (2013, p. 35) found that for Australian academics involved in their project 'it was clear that the language of reference points in relation to forming academic judgements was not familiar to the majority of participants'. Internationally, efforts to align national standards and the development of

qualifications frameworks, Subject Benchmark Statements and Professional standards, learning outcomes, and assessment criteria as described above, have all been based on the belief that previously elusive standards could be made more explicit and, thereby, play an important part in guaranteeing consistency of standards (Universities UK and others, 2010). However such explicit standards promise more than they can deliver (Hawe, 2002) and O'Donovan, Price and Rust (2004) stress the pointlessness of trying to define standards precisely. Research has consistently demonstrated the limitations, even futility (Sadler, 2014), of attempting to make essentially tacit, interpretive knowledge explicit through written expression. For example, Moss and Shutz (2001) argue that such codified standards hide complexity and can mask diversity. Standards have to be used interpretively, but assessors' understanding of terms differs because of their previous experience (Hawe, 2002). These findings have been reflected in a range subjects and contexts (Brooks, 2012).

In relation to research specifically on external examiners, there are interesting findings regarding how they use information to be able to represent community standards. For example, Ross (2009) argues that examiners are bounded by their social and cultural environment and expectations, and Colley and Silver's (2005) research identifies the importance of personal experience of both standards and quality assurance processes in providing examiners' reference points, with less significance given to formal reference points (also see Hawe 2002; QAA 2005). Colley and Silver found that the most important information for examiners was the assessment guidance and criteria for individual tasks within courses, although it could be argued that this is more likely to represent local rather than wider disciplinary standards.

In the context of this broad research, is it appropriate to assume that a system of

independent reviewers or verifiers drawn from academics within the discipline but outside the institution can apply shared knowledge of academic standards and assure that these are consistent and aligned with national frameworks? The research reported below set out to answer this question. The research aimed to explore how external examiners' standards 'in use' are shaped by their personal assessment histories, involvement in professional and/or disciplinary communities, exposure to student work and local and national reference points.

Methods

Twenty-four experienced examiners in chemistry, history, psychology and nursing were recruited from twenty UK universities of varying size and mission group through open advertisement. These participants comprised six examiners from each subject discipline with examining experience ranging from one to twenty years. The project methodology encompassed two data collection methods employed as part of an extended interview. They are described separately here.

Repertory Grid technique

Researchers worked with examiners individually, using a Repertory Grid (KRG) exercise to facilitate the participants in articulating the constructs they use in distinguishing between pieces of student work. KRG is derived from Kelly's (1991) 'personal construct theory' which stresses the active role individuals take in 'construing'; that is making sense of and interpreting events and experiences. KRG aims to capture the dimensions and structure of this personal meaning through an ordered exercise where the participant verbalises the constructs they use in identifying the similarities and differences between people or artefacts. The tacit nature of standards used in assessment means that they are not easily accessible for simple expression by examiners, for example in an interview. Therefore, the KRG method

was selected for its ability to elicit standards 'in use'. Various other studies have used a KRG exercise to undertake research in educational assessment (Johnson & Nadas 2012) because of its ability to elicit how expert examiners construe abstract demands, a key aim of this study.

A week before the interview, examiners were sent five assignments, typical of assessment in their discipline. The assignments were selected because they had been marked as borderline 2.i/ 2.ii (merit/distinction in Australian terms). In all but chemistry, the examiners were also sent a set of assessment criteria for the assignment. Contextual information, such as year and place of study, previously awarded marks and weighting of module, was not provided. In advance of the interview, examiners were asked to read and make notes on the assignments as though marking them.

During the exercise, interviewees were presented with a combination of three out of the five assignments and were asked to identify how two of them were the same but differed from the third. Examiners were then asked to describe the quality in the similar assignments and the contrasting quality in the dissimilar assignment. KRG analysis assumes that these qualities describe the constructs that the examiner uses to think about student work. For example, an examiner stated that two of the three assignments were '*well written with a good academic style'* whereas the dissimilar assignment '*uses colloquial language'*. This reveals that academic style and formal register/language is a characteristic she notices in deciding the quality of student work. The 'opposite' pole is important because it helps identify the examiner's construct more clearly. For example, whilst one examiner positioned *clear academic tone* in contrast to *weak evaluation*, another examiner in the same discipline used a similar construct regarding academic style but considered that the contrast was more to do with appropriate language than evaluative skills (*casual and unscientific language, too story like)*. (See figure 1 for an example of a completed grid)

This process was repeated until all possible trios were exhausted, that is ten times in

total, or until time ran out. In this way, the examiners generated constructs based on an in-the-moment evaluation of actual student work. These self-generated constructs are considered to reflect the personal assessment criteria that the examiners use in discriminating between student work. Examiners were then asked to rank each assignment against these personal criteria and provide an overall grade for each piece. As the grading was not an exacting exercise, analysis concentrated on the examiners' reports of the relative worth of the five assignments when compared with each other, rather than the absolute grade given. The examiners were also asked to rank the constructs they had generated in terms of the

Construct (at 1) (pair of scripts)	**Script (rank 1 to 5)**					**Opposite Construct** (at 5) (single script)	**Priority**
	A	**B**	**C**	**D**	**E**		
Argument excellent	1	2	5	4	3	Argument adequate	1
Less depth and detail of knowledge	4	5	1	1	5	Broad and detailed range of knowledge	1
Expression less fluid	5	2	3	2	1	Well written, rhetorically sophisticated	7
Hardly engages with historiography at all	3	5	2	1	5	Engages well with the historiography	4
Keeps a logical and analytical structure all the way through	1	2	2	3	5	Loose structure	5
Explicitly and critically answers the question	1	2	5	5	1	Not always focused on answering the question	3
Journalistic register	5	4	1	2	4	Academic register	6
Grade (hi, mid, low 3rd, 2:2, 2:1, 1st)**:**	1st	1st	Low 2.1	59/60	1st		

Figure 1. Example of a completed grid displaying constructs elicited.

importance to them in assessing work. The lists of constructs generated by the examiners were scrutinised independently by members of the research team to identify shared meanings across examiners on the basis of the language they used to describe their constructs.

Social world mapping

The second part of the interview focused on the standards external examiners hold and where these come from and led to the construction of a social world map (modified from Clarke, 2005) depicting what they believe to be the provenance of the standards they use as first markers and/or as external examiners. The maps were created in conversation with the researcher and consisted of an A2 sheet. They placed 'elements' (post-it notes) on the maps and organised these around a core and periphery according to how strongly they perceived them to influence their standards. 'Elements' could be people, artefacts, experiences or organisations. Two colours were used to distinguish 'elements' that examiners identified spontaneously from 'elements' emerging in response to specific questions. A third colour was used for pre-completed sources of standards which examiners were invited to add to their maps if they considered them relevant.

The conversations were audio-recorded, transcribed and analysed using a thematic qualitative analysis. The purpose of the map was to discover the provenance of the constructs generated during the KRG by inquiring into the social worlds in which the constructs resided. Discussing membership of different social worlds allowed examiners to describe commitments to these worlds and the ways in which they felt they needed to fulfil them. It also revealed clashes in commitments, identifying examiner awareness of conflicts and how they tried to resolve them.

Findings

The findings are categorised into three sections: standards in use, perceived provenance (location) of standards, and standards as used in external examining. In keeping with a qualitative research design, we have refrained from making statistical inferences from the relatively small number of participants and the non-exacting data collection methods. Rather, analysis concentrates on disclosing examiners' positions with regards to the standards they hold, including how they apply them and from where they are derived (location).

Standards in use

The 24 examiners generated 37 constructs between them with a spread of between three and 10 per examiner and a mean of 7.4. The constructs elicited by KRG were classified as 'global' (33), referring to disciplinary knowledge and academic qualities and 'surface' (4), referring to more generic and technical qualities such as grammar, register, and citation.

Neither the number of constructs elicited from individual examiners nor the overall number of different constructs generated by the group of examiners differed by subject discipline, with each discipline producing between 15 and 18 different constructs. The time constraint built into the KRG method necessarily limited the number of constructs that could be elicited. When asked if there were further criteria that they used in judging student work, eleven participants offered additional constructs. The total number of new constructs mentioned in response to this question totalled four, as many had already been elicited from other examiners during the KRG exercise. Therefore we have some confidence that the method elicited a good picture of the aspects noticed by these examiners in judging student work.

Few clear patterns emerged across the four disciplines. Even when a construct was identified by at least one examiner in each subject discipline, this did not necessarily indicate

strong inter-examiner agreement because of the lack of agreement between examiners within each discipline. For example, *structure and organisation* was found in all subject areas, but whilst five historians used it, only one chemist and one nurse did so. The greatest commonality of constructs across disciplines emerged amongst surface criteria.

There was relatively little sharing of constructs within disciplines. In one discipline a third of the constructs were elicited from only one examiner and only two constructs were elicited from all 6 examiners within a discipline. Seventeen constructs were generated by at least four examiners within a subject area and it was found that the individual ranking of assignments in these 17 cases (that is 1 = a match to the construct and 5 = a match to the opposite construct) varied considerably. There were only nine incidences out of a potential 85 opportunities that all examiners within a subject gave an assignment roughly the same assessment in relation to a specific construct (within two scores) and only two examples where all the examiners awarded the same score. There were 42 instances where examiners rated the five different essays from 1 to 5; that is as both exhibiting the construct and exhibiting the opposite. These results open up the question of how much shared language represents shared interpretation. The examiners used similar language to describe apparently different characteristics or held a different perception of what quality means in relation to the various criteria. This variation in meaning appeared to lead them to rank assignments differently along the same constructs, resulting in manifestly different standards underpinning their judgement.

When the examiners were asked to rank their constructs in order of importance for marking student work, surface constructs were typically ranked as less important than global constructs. There was no other pattern in how the examiners ranked the different constructs that they used with many shared constructs ranked differently by examiners in the same subject area. For example, in the constructs which were largely shared by the historians such

as *structure*, *historiography and academic style*, the rankings ranged between 1 and 5, 2 and 5 and 1 and 10 respectively.

Examiners' overall judgement of the quality of student work, as evidenced by how they graded the assignments, revealed little inter-examiner agreement. Only one of the twenty pieces was assigned the same rank (highest or joint highest) by all six examiners in that discipline. All other 23 assignments were given grades that 'ranked' them against the other assignments in at least three different positions (i.e. best, second best, etc.). Nine of the 20 assignments were ranked both best (or joint best) and worst (or joint worst) by different examiners. This variation in assessment of the work did not appear to be the result of selecting borderline pieces for the exercise where a few marks' variation might make a significant difference to the individual ranking. Instead, the grades offered typically ranged across two to three grade bands. Analysis of the individual construct score indicates that even where the overall judgements about an assignment were similar, examiners frequently made different judgements about the strengths and weaknesses of particular aspects of the work.

In some ways, this inconsistency between the relative overall worth of different assignments /scripts is not surprising given the findings above vis-à-vis the lack of consensus regarding the choice of constructs used to judge the pieces and, where there was construct consensus, the apparent variation in meaning assigned to them. It is worth considering whether some of the examiners may have presumed that we had selected assignments from a range of grade bands and this presumption became part of the context they were working in. Thus they were seeking difference. This is important in considering how external reviewers may be influenced by contextual information such as the grades or grade bands awarded by internal markers.

In summary, the KRG exercise indicates that academic standards, as demonstrated by a sample of experienced external examiners appear to be held by individuals as differentiated

personal constructs. This means that, in the absence of contextual information such as first markers' grades or grade bands, examiners make different assessments of the absolute and relative quality of student work. They use a range of different constructs to discriminate between student performances, they value the constructs differently and they interpret individual constructs sufficiently differently to make manifestly different judgements regarding the quality of student work. The implications of these KRG findings for external examining are explored in the discussion below.

Location of standards

Initial analysis of all 24 maps found that most of the elements on the maps could be categorised within one of two groups. One group comprised explicit standards' documents, such as 'assessment criteria' or 'national benchmarks'. The other group comprised a range of elements relating to personal values or past experiences, including details such as 'school attended' and 'early career mentoring'. In addition, some examiners selected elements relating to ongoing experiences with the potential to shape standards more directly, including moderation and external examining. 'Student work' does not fit in either of the above categories, but appeared on a few maps.

The two main categories identified represent two contrasting ways of conceiving standards--as residing outside the examiner, in explicit documents, or as located within the individual examiner and built up over time through experience. Revisiting the interview transcripts and maps to develop a more nuanced picture, we found that most examiners switched fluidly back and forth between describing standards as internalised or external. A few, however, were adamant that standards should be located in documents because they felt this was most fair to students; for example they felt there was a contract with students to mark according to the assessment criteria. In general, examiners talked about processes that

helped them calibrate standards early in their careers but most of them no longer feel the need to engage in them.

Some examiners were more reflexive about the provenance of their standards and their practices than others; many commented in the interviews that there were few opportunities to reflect on the provenance of their standards or how their standards aligned with those held within the broader disciplinary community.

Standards in the context of external examining

An early surprising observation during the interviews was that some examiners do not see a place for their own standards in the external examining process. Some interviewees could see no connection between the activity of the KRG exercise and the task set for the external examiner; they saw their role strictly as being defined by the institution that had employed them as external examiner. Therefore, in relation to what the external examining system entails and what and whose standards should be used in the examining process, a number of often contrasting views could be observed in the data.

The different viewpoints of the examiners were categorised in two ways: the first was the extent to which they understood their role to be safeguarding discipline **standards** or to be safeguarding assessment **procedures**. The second category was the extent to which they drew on the stated standards of the examined institution as opposed to drawing on wider disciplinary standards. In other words, the first category is about what the role entails in relation to standards, the second is about whose standards are being used in the external examining process. In emphasising these different viewpoints, examiners perceived their roles variously with more or less concern to reflect explicit national standards as set out in qualification frameworks and threshold learning outcomes. Indeed, a significant group perceived that their role was to check whether assessment procedures are followed, and only

in relation to the stated standards of the awarding institution. External examiners adopting this role did not consider standards brought from outside the institution to be relevant as institutions have the authority to set their own standards. Furthermore, those examiners who drew on wider standards appeared to assume that their personal standards represent the national standards for their discipline although there is some evidence from the KRG findings that these vary between examiners. Overall, the research found that examiners can hold very different conceptions of the examining role in relation to standards.

Discussion

What do these findings tell us about the potential for external examining as a process for verifying academic standards? The respondents were experienced academics, selected for their expertise in disciplinary standards. Yet they exhibited very diverse judgements regarding the aspects of student work which they paid attention to, their judgement about those aspects and their overall ranking of different assignments. These results should not be taken as a criticism of the examiners, but as a reflection of the difficulties in the conceptual basis for external examining and, more generally, in assessing university level work consistently and fairly. Student work is complex and unpredictable, there are often no correct answers and considerable latitude exists in how learning can be demonstrated.

The process of external examining assumes that individuals are able to draw upon a shared knowledge of standards. It assumes that their experience and expertise in the discipline and in assessment enables them to make consistent and reliable judgements about the standards in another institution drawing on local or national reference points such as threshold outcomes or qualification frameworks. Yet the KRG findings suggest that such explicit reference points are insufficient, on their own, to enable external examiners to deliver consistent judgement. They are insufficient because a significant number of assessors at this

level do not appear to draw on such reference points in the first place and the meaning of them is interpreted differently contributing to manifestly different appraisals of student work. Therefore a key finding of this study is that external review without calibration of standards cannot serve the purpose of assuring comparability and consistency of standards. Indeed, we have little empirical evidence yet of the potential of community processes to provide effective calibration but this study suggests that, without it, external reviewers may well be applying a personal, rather than a wider discipline-based, interpretation of standards.

In addition, when examiners consciously use explicit reference points such as sets of criteria or benchmarks in making their judgement, they appear to believe that they are interpreting disciplinary standards in a consistent way, unaware of the personalised meanings involved. Similarly where the examiners relied entirely on their internalised standards, they appeared to believe that early career calibration was sufficient to ensure their standards were aligned with others in their disciplinary communities. If examiners are not aware that they hold a personal interpretation of standards, they are unlikely to see the need to engage in on-going calibration processes that help to ensure shared and continuing understanding of such standards amongst disciplinary communities. One explanation for examiners failing to value calibration activities may be that such processes rarely take place in meaningful ways.

A particularly important finding is the manifest variation in judgement in relation to individual criteria. Whilst unreliability in academics' assessment of student work is well-documented, there is little prior data indicating how much specific criteria are interpreted differently although similar evidence was found by Grainger, Purnell and Zipf (2008). This is worthy of further investigation given the emphasis placed on analytical criteria in many qualification frameworks, lists of threshold outcomes and professional standards.

Overall, this study suggests that the adoption of external examining without significant processes to calibrate individuals' standards against negotiated disciplinary norms

will not function to verify standards. External reviewers may act to give the impression of some form of external checking but that checking is likely to be against an individual's own 'standards framework' (Bloxham et al. 2011) and that is only where they are provided with student work uncontaminated by markers' grades, comments or knowledge of the sampling criteria for the work they review. Examiners' responses to the KRG exercise and their reluctance to give grades because they did not have sufficient knowledge of the context reinforces the view that marking is a situated activity and that judgements cannot easily be made in the absence of other 'referencing' information (e.g. grades given, sample examined, knowledge of students' backgrounds, the teaching they received, what tutors expected). If grades are present on the work scrutinised, they are likely to act as the primary reference point in deciding whether standards are appropriate. We would argue that access to this information is a key factor in explaining the extremely high proportion of grading decisions with which external reviewers agree[3] in comparison with the huge diversity in judgement found when they worked with unmarked assignments in this experimental study.

In addition, if external examiners are to provide a cross institutional function in verifying 'national' or 'disciplinary standards', this study suggests that examiners need to understand the importance of using external reference points to inform their internalised grasp of standards. However, in keeping with the sentiments of the last paragraph, these reference points must be understood to provide limited guidance unless they have been subject to sufficient community processes to develop shared meaning.

Conclusions

The general aim of this paper is to draw on the findings of research into how academic standards are conceived, constructed, and applied by external examiners in the UK

[3] See for example the reports from the Australian Go8 Quality Verification System listed on its webpage: http://sydney.edu.au/ab/qvs (accessed 24th October 2013).

with a view to informing the development of external examining and verification methods in other university systems (and indeed in the UK). The study did not seek to problematize the full range of activities that are part of the external examiner role, exploring aspects directly related to the question of the application of standards.

Taken together, the findings raise a number of wider concerns for the development of external examining methods. The research suggests that a UK approach to examining should not be adopted without adaptation. Firstly, it is clearly important to clarify how external reviewers should conceive of their role with regard to safeguarding standards (other responsibilities of examiners are outside the scope of this research) and official guidance should illuminate this. However, if part of the *raison d'etre* of an external examining system is to maintain some sense of national threshold standards, and assessment tasks and student work are the key output measures of those standards, then stakeholders should not be satisfied with a role which is essentially about checking assessment procedures. It is important to develop a role which is fit for the designated purpose and a gradual slide to safeguarding procedures in the UK, as suggested by our examiners, is unlikely to fulfil the aims for external examining discussed in the introduction to this paper. Furthermore, it is unlikely to obtain fairness for students if the emphasis in not on safeguarding standards.

Secondly, acquiring institutional/disciplinary consistency in standards is difficult and dynamic; examiners need formal opportunities to calibrate standards on a regular basis. Therefore review processes that build this into their approach are more likely to make a greater contribution to securing standards. To develop a shared understanding of standards, disciplinary associations, national organisations and institutions should provide examiners with opportunities to engage in a range of activities. These activities should include processes

for reviewers to calibrate their standards within their discipline communities at national level and to align with available reference points. There are good examples of this in practice (Watty et al). These processes should be underpinned by a recognition of the limitations of explicit standards and their relationship to tacit understandings. Furthermore, institutional processes should offer opportunities for examiners to reflect on the provenance of the standards they use; not with the purpose of eliminating personal influences, but rather to raise awareness of them such that examiners can endeavour to resolve inconsistencies between their personal standards framework and national standards. As mentioned earlier, such processes can have a positive impact not only on comparability of standards but also in strengthening discipline communities and increasing professional development (Barrie et al., 2014).

Thirdly, a matter for further consideration is the extent to which external examiners are provided with information regarding grades, grade bands or samples in work that is scrutinised. If external examiners' central role is related to safeguarding standards, this research suggests that they will struggle to exercise independent judgement if influenced (however unconsciously) by knowledge of the initial grades and knowledge of the student. Additionally, there is some evidence in this research that examiners are concerned that negative appraisals may affect later relationships and employment opportunities. Therefore, having an anonymous external review system as set out in Krause et al. (2014) where universities do not know who is verifying their work and reviewers do not know who they are verifying warrants consideration.

Finally, any system needs to recognise the limitations of explicit statements of standards which have little power to assure consistency on their own. They can provide

reference points for calibration discussion, but their value is only really obtained through the development of shared meaning.

Sponsorship The research reported in this paper was jointly sponsored by the UK Quality Assurance Agency and the UK Higher Education Academy.

References

Australian Government. (2013). *Draft standards for course design and learning outcomes*. Canberra.

Barrie, S., Hughes, C., Crisp, G. & Bennison, A. (2014). *Assessing and assuring Australian graduate learning outcomes: principles and practices within and across disciplines, Final Report*. Sydney Australia: Office for Learning and Teaching

Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: The role of assessment criteria in UK higher education grading practices. *Studies in Higher Education*, 36(6), 655-616.

Bloxham, S & Boyd, P., (2012). Accountability in grading student work: Securing academic standards in a 21st century quality assurance context. *British Educational Research Journal*, 38(4), 615-634

Bloxham, S & Price, M (2013). External examining: fit for purpose? *Studies in Higher Education*, 1-13. doi: 10.1080/03075079.2013.823931.

Bradley, D., Noonan, P., Nugent, H., & Scales, B. (2008). *Review of Australian higher education: Final report* (Bradley Report). Canberra: Department of Education, Employment and Workplace Relations.

Brooks, V. (2012). Marking as judgement. *Research Papers in Education*, 27(1), 63-80.

Colley, H. & Silver, H. (2005). *External examiners and the benchmarking of standards.* York: Higher Education Academy.

Clarke, A. (2005). *Situational analysis: Grounded theory after the postmodern turn*. Thousand Oaks, CA: Sage.

Deane, L. & Krause, K. (2012). *Towards a Learning Standards framework, Learning and teaching standards (LaTS) Project: Peer review and moderation in the disciplines.* http://www.uws.edu.au/__data/assets/pdf_file/0010/398620/Learning_Stds_Framewk_Final_Dec_2012.pdf Accessed Nov 1st 2013

Dill, D. D., & Beerkens, M. (2012). Designing the framework conditions for assuring academic standards: Lessons learned about professional, market, and government regulation of academic quality. *Higher Education*, 65(3), 341-357.

Grainger, P., Purnell, K. & Zipf, R., (2008). Judging quality through substantive conversations between markers, *Assessment and Evaluation in Higher Education*, 33(2), 133–42.

Hawe, E. (2002). Assessment in a pre-service teacher education programme: the rhetoric and the practice of standards-based assessment. *Asia Pacific Journal of Teacher Education*, 30, 93-106.

Johnson, M. & Nadas, R. (2012) A review of the uses of the Kelly's Repertory Grid method in educational assessment and comparability research studies, *Educational Research and Evaluation: An International Journal on Theory and Practice*, 18 (5) Published online 20th June 2012. DOI:10.1080/13803611.2012.689715

Kelly, G.A. (1991). *The psychology of personal constructs: Volume 1: A theory of*

personality. London, UK: Routledge. (Original work published 1955).

Krause, K., Scott, G., Aubin, K., Alexander, H.,Angelo, T., Campbell, S., Carroll, M., Deane, E., Nulty, D., Pattison, P., Probert, B., Sachs, J.,Solomonides, I., Vaughan, S. (2013). *Assuring final year subject and program achievement standards through inter-university peer review and moderation*. Available online: www.uws.edu.au/latstandards.

Moss, P.A. & Schutz, A. (2001). Educational Standards, Assessment and the search for consensus. *American Educational Research Journal*, 38(1), 37-70.

O'Donovan, B., Price, M. & Rust, C. (2004). Know what I mean? Enhancing student understanding of assessment standards and criteria. *Teaching in Higher Education*, 9(3), 325-335.

O'Hagan, S.R. & Wigglesworth, G (2014). Who's marking my essay? The assessment of non-native-speaker and native-speaker undergraduate essays in an Australian higher education context, *Studies in Higher Education*, Published online: 08 Apr 2014, http://dx.doi.org/10.1080/03075079.2014.896890

Price, M. (2005). Assessment standards: The role of communities of practice and the scholarship of assessment. *Assessment & Evaluation in Higher Education*, 30(3), 215-230

QAA (2005). *Outcomes from institutional audits: external examiners and their reports*. Gloucester: Quality Assurance Agency.

QAA (2011). *UK Quality Code for Higher Education, Part B: Assuring and enhancing academic quality, Chapter B7: External Examining*. Gloucester: Quality Assurance Agency.

Ross, V. (2009). External music examiners: micro-macro tasks in quality assurance practices. *Music Education Research*, 11(4), 473-484.

Sadler, D. R. (2007). Perils in the meticulous specification of goals and assessment criteria. *Assessment in Education: Principles, Policy & Practice*, 14(3), 387-392. doi: 10.1080/09695940701592097

Sadler, D.R, (2014). The futility of attempting to codify academic achievement Standards, *Higher Education*. 67, 273–288 DOI 10.1007/s10734-013-9649-1

Shay, S.B. (2004). The Assessment of Complex Performance: A Socially Situated Interpretive Act. *Harvard Educational Review* 74(3), 307-329.

Trowler, P. (2009). Beyond epistemological essentialism: Academic tribes in the 21st century. In C. Kreber (Ed.), *The university and its disciplines: Teaching and learning within and beyond disciplinary boundaries* (181-196). New York: Taylor & Francis.

Universities UK, Guild of Higher Education, QAA (2010). *Review of External Examining Arrangements in the UK: A discussion paper from Universities UK, Guild HE and the Quality Assurance Agency for Higher Education*. London: Universities UK.

Watty, K., Freeman, M., Howieson, B., Hancock, P., O'Connell, b., de Lange, P & Abraham, A. (2013) Social moderation, assessment and assuring standards for accounting graduates, *Assessment and Evaluation in Higher Education*. Published online 11 Nov 2013.

Wenger, E. (1998) Communities of practice: learning, meaning and identity. Cambridge: Cambridge University Press.

Yorke, M. (2009). Assessment for career and citizenship. In C. Kreber (Ed.), *The university and its disciplines* (221-230). London: Taylor & Francis.