[Un] reliable assessment: A case study

YVONNE REED, STELLA GRANVILLE, HILARY JANKS, PINKY MAKOE, PIPPA STEIN AND SUSAN VAN ZYL WITH MICHAEL SAMUEL[1]

Yvonne Reed, Stella Granville, Hilary Janks, Pinky Makoe, Pippa Stein and Susan van Zyl all teach in Applied English Language Studies at the University of the Witwatersand.

***YVONNE REED'S** research is in the areas of assessment and in the design and delivery of distance learning materials.*

***STELLA GRANVILLE'S** current research interest is in the experiences of teaching and learning of first year student participants in foundation courses.*

***HILARY JANKS'S** main research interests are language education and critical literacy.*

***PINKY MAKOE'S** current research includes early childhood language and literacy learning and language policy.*

***PIPPA STEIN'S** research focus is on literacies, multimodality and pedagogy.*

***SUSAN VAN ZYL** has just been seconded to the Wits Institute for Social and Economic Research and to Psychology where she will pursue her multidisciplinary research interests which include an interest in the relationship between forms of knowledge and forms of writing.*

***MICHAEL SAMUEL** is Chair of Educational Research at the University of Durban-Westville. His main interests are in language education and teacher professional development.*

Abstract

The drive towards quality assurance at South African universities, with 'consistency' of approach being one of its key features, has profound implications for assessment policies and practices in relation to equity. In this article we present a case study discussion of an investigation we undertook, as a department, into certain anomalies which arose in the assessment of a particular group of post-graduate students' research reports. We were puzzled by the variability in the marks awarded by three different markers of the same reports and set out to investigate what factors were producing this 'inter-marker [un]reliability'. Through a content and discourse analysis of the different assessors' written reports, we uncovered the implicit assessment categories and criteria which assessors were working with in their assessments. We discovered shared categories and criteria, as well as differences in how these were weighted. In the interests of equity and increased inter-marker reliability, we have developed a set of banded criteria on generic features of the research report which we intend to

[1] Michael Samuel, University of Durban-Westville, is the external examiner referred to in this article. Where his comments form part of the data we have worked with in the main body of the text, he is referred to simply as the external examiner. We asked him to respond to the first draft of this article and we have included those responses in footnotes under his name.

trial. We also surfaced two unresolved issues: the use of language and the role of the writer's 'voice' in the research report. As a result of this investigation, we argue that the 'consistency' of assessment within and across universities aspired to by quality assurers (such as the HEQC in the South African context) is difficult to achieve and much still depends on professional judgement, intellectual position and personal taste.

In this article, six staff in Applied English Language Studies (AELS) at the University of the Witwatersrand, analyse the assessment of Honours research reports in one particular year.[1] What puzzled the staff was the variability in the marks awarded by three different markers of the same reports. The article is divided into five sections:

- an introduction which locates the case within assessment in Higher Education;
- a description of the case itself and the method of enquiry;
- data analysis which enabled the production of agreed assessment criteria that could be banded in categories;
- data analysis which produced disagreement particularly on the issue of 'voice';
- implications for assessment practice.

Introduction

In 1999 the University of Western Australia's on-line newsletter, *Issues of Teaching and Learning*, noted that the assessment of dissertations and theses is an area in which there has been limited research and discussion (http://www.csd.uwa.edu.au/newsletter/issue0899). Brown, Bull and Pendlebury (1997) writing on project work in higher education which culminates in a research report or dissertation, note that there has not yet been substantial research on such project work or its assessment (121). They suggest that three common difficulties in the assessment of projects are 'variations between assessors, variations in the difficulty of projects and the variations in students' and supervisors' contribution to the project report' (127). The focus of this article is on the first of these difficulties: variations among assessors. We foreground issues pertaining to variation in relation to (1) [un]shared assessment criteria, (2) weighting of these criteria and (3) judgement based on differing intellectual 'positions'.

At a time when South African universities are expected to pay more attention than ever before to quality assurance, with 'consistency' of approach being one of its key features (Gipps, 1994; Smout, 2002), we argue that while the specification of explicit assessment criteria can go some way towards providing students and their assessors with guidelines on project/task requirements, such criteria do not guarantee reliability of assessment and in the end much still depends on professional judgement. We have used Gipps' (1994) definition of reliability:

> Reliability is concerned with the accuracy with which the test measures the skill or attainment it is designed to measure. The underlying reliability questions are: would an assessment produce the same or similar score on two occasions or if given by two assessors? Reliability therefore relates to consistency of pupil performance and

[2] We have chosen not to disclose the year to protect the identities of students involved and because we believe that the issues raised do not pertain only to a particular year or group of students. We believe that the particular case that we have considered can be generalised to other years and groups that we have taught, and we hope that readers will recognise some of their own practices in this discussion of our work.

> consistency in assessing that performance: which we may term replicability and comparability (67).

In response to it, this article raises questions about the difficulties in achieving inter-marker reliability even when using marking bands and grade descriptions. We argue that values and judgements concerning intellectual position and interests, as well as personal taste or professional judgement, are deeply implicated in how marks are awarded.

Presenting and investigating the case

Context and research puzzle

In Applied English Language Studies (AELS), the Honours course consists of five 'papers', each of which counts 20% of the final coursework and examination total. The research report is one of these five papers and thus counts 20% of the overall total of coursework and examination marks. Students are expected to complete a research report on a subject of their choice. They prepare for their research project by participating in a semester-long research methods course in the first half of the academic year. The generic features of the research report are outlined in a class handout which includes information about what is expected in the abstract, introduction and rationale, literature review, research methods, data description, data analysis as well as an analysis of implications of the research findings. Students are expected to understand the importance of acknowledging sources and of accurate referencing. They are also expected to use standard English and to pay attention to the overall presentation of their work. In the past, they have not been provided with explicit assessment criteria: we have relied on the process of supervision to make students aware of what is expected of them.

The usual practice at the University of the Witwatersrand is for the report to be assessed by the supervisor, who assigns a mark and writes an assessment report. The marks and the assessment reports are then sent to the external examiner who is required to moderate every research report. In the case under discussion, the course co-ordinator expressed concern about the marks awarded by the supervisors as she felt that the mismatch between some students' final marks for other papers and the mark for their research report was problematic. It was agreed to ask a member of staff who had not supervised any of the students to assess the reports. She was not given access to the mark awarded by the supervisor. She assigned a mark to each research report and provided a written explanation for this mark. This second marking further confused the issue as there was little match between this set of marks and the first set of marks. On the basis of this evidence of inter-marker unreliability, we agreed to average the two marks and to send the external examiner all three sets of marks: the supervisor's mark (A), the second marker's mark (B) and the average of the supervisor's and second marker's mark (C). An explanation of the process and of the reasons for the decisions that had been taken, accompanied the marks.

The external examiner 'read and re-read'[2] each research report and then assigned a new, different mark to each report, which we refer to as the external examiner's mark (D). This mark, the fourth to be allocated in this process, did not have any discernible patterned relation to any of the other three. Table 1 provides an overview of the four marks allocated to each research report. The external examiner gave us the option of averaging his mark with the previous

[3] Quote from the external examiner's report.

average, but suggested that he would prefer his mark to stand. We deferred to his preference and his mark was the mark finally awarded to students. The table shows inter-marker reliability in only two instances. There is no pattern of the second assessor's mark being consistently higher or lower than that of the supervisors: four of her marks are higher, two are lower, and two concur. There is also no pattern of the external examiner agreeing with any of the marks: in two cases, his mark is closer to the average mark, in two cases he agrees with the supervisor, in two cases, the marks all concur, and in the remaining two cases, he is closer to the second marker.

Table 1: Overview of the four marks allocated to each research report

Student	A Supervisor's mark	B Second marker's mark	C Average: supervisor and second marker	D External examiner's mark
1	63	38	50.5	48
2	65	85	75	68
3	72	76	74	74
4	55	63	59	54
5	57	35	46	42
6	68	68	68	68
7	60/54[3]	70	61	70
8	66	67	66.5	68

Investigating the case

The process of investigation involved a number of steps. We began with the hypothesis that the assessors were all working with different implicit assessment categories and criteria, and we set out to make these explicit. The first step involved AELS staff in performing a content and discourse analysis of all three assessors' written reports in order to 'surface' the implicit assessment categories and criteria evident in these reports. This process was a form of 'grounded analysis' in that we were looking for categories and criteria to emerge from the data (i.e. the assessors' written reports). We each analysed a full set of reports on a student, other than the one(s) we had supervised. Our task was to select data, in the form of 'useful quotes' from each report (supervisor's, second assessor's and external examiner's), which provided evidence of each assessor's use of implicit or explicit categories, criteria or assumptions in the assessment of the research reports.

In the second step of this process, we labelled sheets of newsprint with headings taken from our analysis of the assessors' reports. These headings categorised generic features of the research report as they emerged from the data. Under these category headings we explored what we understood by each of them and by their related specific assessment criteria. This was a messy process: we started with the assessors' reports but felt too constrained by them and moved into general brainstorming of any ideas we individually and collectively associated with these categories and criteria.

We had begun our investigation with the hypothesis that the assessors were all working with different assessment criteria. We quickly discovered, much to our surprise, that both at the level of form and content, all three assessors were operating with shared criteria. These related to the

[4] This student had two supervisors each of whom awarded a separate mark.

generic features of the research report, such as the literature review, method and data analysis as well as to language use and presentation.

Further investigation revealed that the mark discrepancies related to the weighting each assessor gave to these shared criteria. Two glaring examples of different weighting related to the literature review and the use of language. Whereas the external examiner attached more importance to the development and use of the literature review and the use of literature than did some of the AELS staff, the second marker gave greater importance than other markers to the use of language.

Importantly, we discovered that there was a major area of disagreement around the role and value of the writer's 'voice' in the research report. We refer to the different positions on voice as 'the voice cline' and define it as the extent to which the student as person is present in the research report and the extent to which the student as 'teacher/practitioner' is present in a specific context. Discussion around this issue, presented in more detail later in this article, was both extensive and heated. Positions ranged from those of some staff who thought voice was not an issue at all, to that of the external examiner for whom it was clearly a major issue. In addition, we found that there was some misunderstanding as to whether Honours level research is expected to make a new contribution to knowledge and as to what constitutes new.

As a consequence of this investigation we decided to take action on our assessment practices in two ways: at the practical level, we agreed to work out a set of categories and within these, banded criteria or 'grade descriptions' for generic features of the research report based on our shared criteria, with the aim of giving assessors and students more explicit guidelines on what was expected of them. We also agreed to institute, as a regular practice in the assessment of research reports, a second marker, who would mark the reports 'blind' and to give the external examiner the marks of both the supervisor and the second marker. On a more theoretical level, we decided to explore the issue of voice, on which there is still no shared position.

Shared assessment categories and criteria

Despite our reservations, both theoretical and practical, in relation to the establishment of explicit criteria for research reports, we have produced criterial statements, in four categories, which we regard as provisional, until we have built up experience in using them. See Table 2. We have limited the number of categories to four in order to facilitate their use as an assessment tool. The categories, which emerged from our analysis of the assessment reports and which we will discuss later, are:

- use of literature in the research report;
- the research process, including the research question, the method, the collection and quality of the data and the data analysis;
- the use of language, control of the discourse and genre, presentation;
- overall impression.

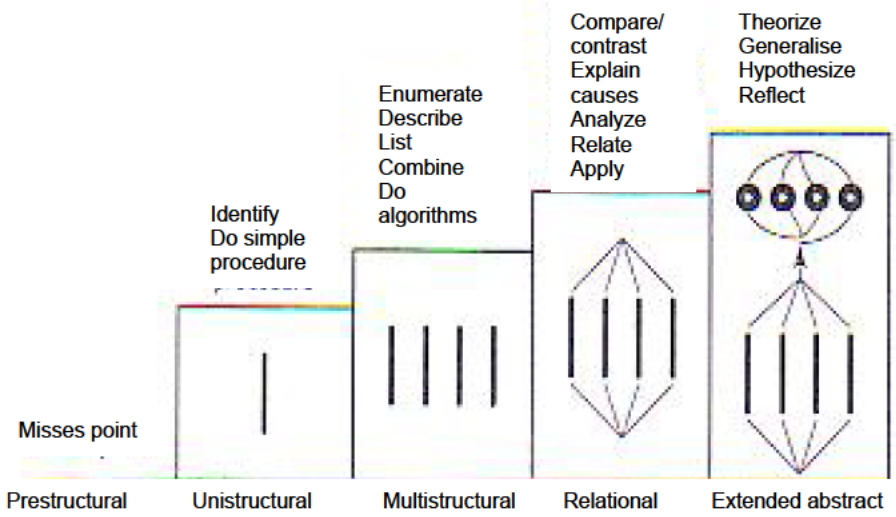
These categories, with the exception of language, have been organised into achievement levels to produce a grade description. This results in a grade band for each category which captures some of the subtleties/nuances which we uncovered in our analysis of the assessment reports. Each grade includes the achievements of the grades below, so, for example, Grade C includes grade D and Grade A includes Grades B, C and D. This is symbolised by the dotted line with an arrow that runs across the top of Table 2.

Table 2: Banded criteria ←----------

Biggs' SOLO taxonomy	**Extended abstract**: Theorise, generalise, hypothesise, reflect		**Relational:** Compare/contrast, explain causes, analyse, relate, apply		**Multi-structural** Enumerate, describe	**Uni-structural** One aspect only	**Pre-structural**
Wits grades	A >84%	A 75-84%	B 70-74%	C 60—69%	D 50-59%	E < 50%	
Use of literature in the research report	Provide a fresh or original synthesis of ideas in the literature Use the research to raise questions about the literature Mastery of the demands and purposes of the genre of a literature review	Able to evaluate the different positions in the literature and to engage with them critically Fluent with literature that is up-to-date and relevant to the project Able to use the literature to provide a nuanced interpretation of the data Ability to apply the literature to new contexts	Synthesise the literature in order to formulate an argument in relation to the research project Able to link the reading to discussion and analysis of the research data	Well-documented, systematic review of the reading that goes beyond reference to foundational texts Includes local and international literature Texts compared and contrasted with links made across them Literature provides the theoretical framework for the project References included into the text with ease	Identifies the appropriate areas of literature that need to be covered and provides a sound descriptive summary of the key foundational texts without seeing their relevance to each other or to the project Not able to incorporate references fluently	Inappropriate choice of texts without a clear focus or helpful organisation of the material Misses the point or shows a poor understanding One-dimensional Overreliance on sources to carry the argument	
The research process: question, method, data and data analysis, ethics	Meta-awareness of the research methods and the possibilities that are generated by the data Able to extrapolate the data to other situations and practices and to recognise its significance Able to relate the findings to the literature and to measure their contribution to knowledge Careful consideration given to ethical issues	Ability to work creatively with research methods in the discipline Fluency with the data; careful selection based on an evaluation of the data; ability to incorporate quotes and details Develop a coherent argument that makes sense of the complexities and contradictions in the data Careful consideration given to ethical issues	Integration of question, method, data collection and analysis Able to use the reading to interpret and explain the data Evidence of triangulation where necessary Careful consideration given to ethical issues	Able to identify the patterns and ruptures in the data and to account for them, using the theory provided by the literature Ability to use the data as evidence to answer the research question Careful consideration given to ethical issues	Appropriate methods chosen to produce viable data Clear explanation of the research method, systematic organisation of the data and an accurate description	Poor links between the question, the methods, the data, the analysis and the literature Problematic question, or method or data collection Problematic data: insufficient, haphazard Problematic data analysis: poor selection, poor organisation and analysis	
Language, discourse, genre and presentation	Mastery of the discourse Sophisticated use of language Complex ideas expressed with both flair and reader awareness Controlled and flexible use of the genre Excellent design and professional presentation ←-------			Ability to use clear, simple, error-free English Unstable use of academic conventions and the language of the discipline Sufficient organisation and coherence for the reader to understand Too reliant on quotes Presentation needs more care		Incorrect use of English, riddled with surface errors Problematic syntax Hard to read Incoherent structure and poor organisation of material Unprofessional presentation of work	
Overall impression	Stands out Memorable	Excellent, well-written, integrated project	The research report is complex, interesting and well written	Sound research Competent use of language	Too descriptive Pedestrian Limited overall or flawed in part	Conceptually weak Lack of organisation Poorly written	

In order to establish the bands, we used Biggs' SOLO Taxonomy (Biggs & Collis, 1982). This provided us with an established and theorised framework for defining students' levels of achievement. Biggs' taxonomy is a tool for helping teachers to structure levels of intellectual understanding, not only for assessment purposes, but also for defining desired learning outcomes in a range of different teaching and learning contexts. 'The SOLO taxonomy provides a systematic way of describing how a learner's performance grows in complexity when mastering many academic tasks' (Biggs, 1999, 37). As can be seen in Biggs' diagram below, the framework works on five levels:

Figure 1: A hierarchy of verbs that may be used to form curriculum objectives



A student performing at the pre-structural level misses the point altogether; at the unistructural level he/she can deal with only one aspect of the problem at issue; at the multistructural level a student can list or identify a number of issues but cannot relate them to one another. Higher order thinking is expressed by the relational and extended abstract levels: at the relational level a student can relate, link and see connections – at the extended abstract level he/she can think meta-cognitively, can theorise, reflect and critique the matter at hand.

What we have done is to adapt and modify this taxonomy for our particular purposes. See Table 2. We agreed that neither of Biggs' lower orders of thinking was passable at a postgraduate level. In order to make finer distinctions at the higher order levels, we have split each of them into two. In this way we distinguish between the student who obtains a 'C' (60-69%) and one who gets a 'B' (70-79%). The 'C 'student is one who can classify and organise data competently and relate it to the literature but without any particular sophistication. The 'B' students can integrate all aspects of the project, including the literature. Distinctive students (75-84%) can theorise and critique both their own work and the literature and can evaluate the significance of what they have done. The 85%+ student can produce fresh ideas arising out of the project and can generate insights beyond its immediate concerns.

Finding the categories to organise into bands was less straightforward than might have been expected. Originally we set out to provide assessment criteria for each aspect of the research report genre. For example, we began with the literature review and quickly realised that it was difficult to isolate this section of the report because of the way literature is used both to frame

research and to analyse the data. As a result, we changed the category to 'use of literature in the research report', which enabled us to distinguish between students who could integrate the literature successfully into the report as a whole and those who could not. This then led us to consider the research process as another crucial category. Biggs' taxonomy was particularly suitable for these two categories because it provided a means for us to identify the different levels of thinking in relation to the arguments, ideas, content and understanding, embedded in both the analysis and use of literature and in the production and analysis of data. When one considers language and form as integral to the production of argument and analysis in the first two categories, it is easier to map it onto Biggs' taxonomy. We agree that language and cognition are intimately bound up and that poor use of language affects students' overall performance. For example, it is not possible to formulate an argument or to provide a nuanced interpretation of the data without control of the language. In addition, we also chose to isolate language as a separate criterion to establish this as an important and visible assessment outcome for students. In isolating language in this way, we experienced difficulties in mapping bands onto the Biggs' taxonomy, so we set up a continuum across the grades, signalling only the difference between a pass, fail and distinction. Because language is implicated in the assessment of the other categories we gave the language category a lower weighting than the others. Making a professional judgement about the mark to award for language is complicated by the fact that many of the students who write research reports in AELS use English as either an additional or a foreign language.

In the case described in this article, weighting, rather than assessment categories and the criteria deriving from these, proved to be the main source of inter-marker variability. The weightings that we have decided to work with for the time being remain contested. We have also realised that if we cannot agree on weighting, different markers will be likely to continue to accord too much importance to one or other of the assessment categories. We have decided to work provisionally with the following weightings and to review them after we have gained some experience in using them. We have agreed on the following: use of literature 30%; research and the research process 40% (given the greatest weight because we see it as the heart of a research report), language as an isolated category 20% and overall impression 10%. Choosing to include overall impression as a separate category, provides an opportunity for the assessor to consider the research report as a whole and to consider its overall impact.

In offering this set of categories and descriptors to our students and to the readers of this research, we are aware of the dangers. Banded criteria are an attempt to make implicit expectations explicit but we recognise that

- it is impossible to fully specify criteria;
- making them too specific might act as a straitjacket for students;
- different markers will nevertheless evaluate a student's performance differently.

Different taste, different intellectual positions, different values, and different professional judgement in relation to each of these criteria are likely to continue to produce inter-marker 'unreliability'.

Discussion of the unshared criteria: The voice cline

Whether students' voices should, as the external examiner's reports suggested, appear in the report caused heated debate among the AELS staff. The section that follows attempts to convey some of the different positions staff took on the question of voice by focussing more on those areas where substantial disagreement occurred, which in turn indicated that some people were

located at either end of what we have called 'the voice cline'. An interesting aspect of this discussion is that while some people felt that they did occupy a definable position (usually at either end of the cline), there were others who felt that they could not locate themselves on a cline at all because issues of 'voice' were for them always specific to the context and nature of a particular research project.

Issues of 'voice', which are usually associated with postmodern and poststructuralist discourses, take a number of forms but in the context of an article concerned with the assessment of research reports, we will focus on those which surface in relation to questions of knowledge and knowledge production. In this context, those who support voice do so from a perspectivist position which links knowledge and truth claims to a particular culture or social interest group, form of life or standpoint and in this reject the idea of universal and independent criteria (Moore & Muller, 1999). In other words this perspective is based on the claim that knowledge cannot be separated from statements about the knowers. Knowledge is, as a result, dissolved into knowing and priority is given to the particularities of what is expressed as a consequence of category membership and identity (Maton, 1998).

This intersection of questions of voice with those of knowledge often relates to the liberatory agendas associated with work in education around antiracism, feminism, sexuality and critical pedagogy, all of which explore ways in which 'marginalised' or 'suppressed' voices can be heard. For this reason those in favour of voice usually also support the 'autobiographical turn' in which personal journals and other expressions of identity and the use of the first person become important parts of how the research text is constituted and represented.

In order to focus the discussion on voice (which first took place in the whole group and then was considered in more detail by the two people responsible for writing up this issue), we began with an analysis of the following comments made by the external examiner in his assessment reports about the absence or presence of voice in our students' research reports.

> Extract 1
> It's disappointing that the writer chose a rather clinical detached form of reporting on what is a complex, rich and fascinating research opportunity. The detached writing resulted in fostering a 'cold approach' to a rich ethnographic insider report. The elements are all there but it lacks conviction. (Extract from external examiner's report on an individual student's research report.)
>
> Extract 2
> I am concerned about the mechanistic summaries of past researchers. The report writers/students seem to have lost their voice and the other researchers are foregrounded. It is the duty of the student to present an argument (theirs) *supported* by other researchers. (Italics in the original text; extract from external examiner's overall report submitted at the end of the moderation procedure.)
>
> Extract 3
> Evidence of a confident student who has mastered the genre of writing academically. (Extract from external examiner's report on an individual student's research report.)

On the basis of the external examiner's notes and reports on individual students, as well as his final assessment report submitted to AELS and the Faculty of Humanities, we identified four meanings of the term 'voice'. These are:

- voice as the expression of intellectual judgement / independent mindedness / argument / intellectual position;
- voice as the expression of personal experience and practitioner knowledge;

- voice as conveying a sense of conviction;
- voice as epistemologically clarifying.

Voice as expression of intellectual judgement/independent mindedness/ argument/intellectual position

This view of voice seems to relate to the common belief that students should present an argument that is in some sense their own, which locates their work within a chosen theoretical position, thereby displaying forms of intellectual judgement. In Extract 2 above, the external examiner suggests that problems related to voice come to the fore in the context of the literature review where our students 'seem to have lost their voice' in relation to 'other researchers'. He relates this absence of personal voice to an absence of confidence, exemplified in the way students take on the theories of other researchers as 'gospel truths'.

There was general agreement that it is correct that these judgements should emerge in the literature review, that they should underpin and orient the research and should be retrieved in the conclusion. It was also agreed that students should be helped to develop the ability to make an intellectual judgement and to read sources, authorities and theorists critically. However, there was a difference of opinion related to the level of study at which this kind of judgement could be expected, let alone required. We agreed that students at Honours level should be able to locate themselves in relation to one set of theories rather than another and be able to articulate why a particular approach was more useful for their own research than another. However, one of us did not believe Honours level students should necessarily be expected to take critical positions in relation to authorities in the field or that there was any problem with treating an authority as an authority. In response it was suggested that the 'stature'of the theorist in question might be a consideration to take into account because it is very hard for an Honours student to critique the complex, canonical work of Freud, Vygotsky or Chomsky, for example. By contrast, they should, even at Honours level, be expected to read secondary sources or more accessible theorists critically.

There was definite opposition from this staff member to the idea that reading critically had anything to do with the question of 'absolute truth' and the implication that a student who does not contest his or her sources is somehow displaying a lack of confidence and critical thought. She referred to this position as a form of 'false egalitarianism' because she thought the implication that the novice's personal opinion is inherently worthwhile is simply 'naïve and misleading'. The opposing view was sympathetic to the idea of encouraging students to contest and challenge authorities and sources regardless of their level of expertise in the light of the history of Bantu Education and the silencing of marginalised and resistant voices during the apartheid era.[4]

[4] The external examiner subsequently confirmed that he agreed with this view and that his comments on the importance of 'voice' could be read as an attempt to infuse a debate about what knowledge is produced, and whose knowledge ultimately surfaces as the output of a (white) (liberal) university education. He adds,

> This is not only a question about giving voice to the previously 'disempowered'. It is about how the future generation of South African are interpreting their role in relation to the numerous 'imported theories' that seem to easily find a marketplace in our university education system. Who are the midwives of such knowledge consumption? We need to ask whether we could confidently list at least 10 South Africa theorists whom we think are being quoted as providing new insights and knowledge. Who are our theory-builders? In the present dominant university system we seem to believe that our theory-builders are those who have demonstrated their ability to mimic the research traditions from the outside (western) (developed) world. Yet we know that these 'gurus' to whom we defer usually develop their theoretical conceptions based on worlds which differ fundamentally from the worlds we as African, South African live in. Where arte the great theoreticians in South Africa? Who are they? Why do we believe them?
>
> (Samuel, personal communication, September, 2002)

Everybody agreed that it was important to develop students' confidence as thinkers and producers of knowledge and that this is central to academic development. However, not everybody thought this was a challenge which should be met by focusing on voice.

Voice as expression of personal experience and practitioner knowledge

Voice as 'personal' experience is understood by us to refer to the personal in what is traditionally associated with personal identity – personal experience of class, race, gender, attitudes, values, ideological orientations and political beliefs. Here opinion was definitely divided between those who supported voice in all or most of the meanings listed above and those who would argue that voice, in the 'personal' sense, should only be invoked on those (rare) occasions when these factors explicitly influence and are logically relevant to the nature of the research project.

In discussion we discovered that this latter position was based on the belief that 'personal experience' and 'practitioner experience' could be distinguished by rational means. In other words, there is no good reason for believing that issues like the influence of class, race and gender and their possible role in the research cannot be reflected upon by researchers in the same way that they reflect on other kinds of intellectual questions or research practices. In this view, there is no reason to assume that because these things are supposedly more 'intimately part of our subjectivity' that they should be expressed on all occasions, and that this expression resolves what is problematic about them. One staff member described the views in support of the invoking of personal voice as 'flavour of the month' opinions that do not stand up to serious scrutiny.

By contrast, staff members working within a broadly constructivist orientation could not subscribe to the view that in reflective practitioner research, ethnography and similar forms of qualitative research, the relationship between 'knowledge' and 'the knower' could any longer be believed to be unproblematic. Issues of identity, culture, history and standpoint are implicated in the knowledge production and representation and must be taken into consideration wherever truth claims are made.

At this point in the discussion, it became clear that those largely in support of the expression of voice in research were supporting it in the case of teacher research in particular, where practitioner experience is considered to be particularly important. In this, staff members were aligning themselves with the view of Freeman (1998, preface) who situates teachers' voices and points of view firmly within teacher-research as an activity that connects the 'doing' of research with the 'questioning' of research. The external examiner's view that questions of identity and conviction that express themselves in voice are fundamental to defining what counts as teacher-based research in particular contexts and historical moments, was on the whole, supported by those involved in teacher education and educational research.[5]

[5] He subsequently added to the discussion with this comment,

> Presenting the personal positionality is not mere ornamentation. It allows the reader to be able to locate the writer of the text, to be able to critique the intended (sometimes implicit) dialogue that she wishes to develop with an audience. The audience is able to creatively and critically locate how they then wish to engage with the author of the text. It allows for continuing a 'creative discursive space' to be established which allows access for readers to enter into the knowledge production process. Research is not about whom to keep outside the space, but about providing a democratic invitation to contribute.
>
> (Samuel, personal communication, September , 2002)

Voice as conveying a sense of conviction

Extract 1 from the external examiner uses the phrases, 'a rather clinical detached form of reporting' and 'a cold approach' to describe how a particular student engaged with her research topic which was ethnographic in style. Written into these comments is an assumption that the ethnographic and participant research traditions require the expression of some form of commitment, 'warmth' or 'passion', from the researcher. In other words, an explicit expression of engagement is one of the criteria by which the writing up of research of this kind is assessed.

This view became controversial when it seemed to suggest that 'thick description' should (always) be 'hot' as distinct from 'cold'. Implicit in this position is the belief in a strong relationship between voice and commitment – that is that 'voice as conviction' is desirable and that practitioner research in the ethnographic tradition must in some sense move beyond detachment into a type of 'advocacy'. At least one person did not support this view on the grounds that an ethnographic researcher is not always an insider and even where this is the case 'warmth' or 'conviction' must be seen as strictly optional. She referred to some of the most famous and detailed ethnographic work (such as Geertz's classic description of Balinese cock fighting which expresses no personal convictions on the matter of cock fighting).[7]

Voice as epistemologically clarifying

This aspect of the voice debate is related to the belief that expressing yourself in your own voice (as a means of capturing your 'subjectivity') is an epistemological virtue and part of what Walkerdine (1997) calls 'the laudable autobiographical turn'. In other words, voice as a means by which your particular identity may be 'confessed' is believed to go some way towards the validation of truth claims that might otherwise be biased. This view is associated with that in *Writing Culture* where it is suggested that 'if ethnographic truths are inherently partial, committed and incomplete, a rigorous sense of partiality can be a source of representational tact' (Clifford and Marcus, 1986, xx).

Where the author or researcher is seen as a site of hidden prejudices and opacities, then an attempt to uncover these prejudices by way of a form of self disclosure is believed to 'reduce' bias and prejudice. We agreed that this aspect of the 'voice' question could be related to a widespread conviction that to express subjectivity and to attempt to articulate it is epistemologically desirable. But we did not all agree that this was in fact the case or that 'confession' was epistemologically helpful. In fact one staff member described this view as confusing and certainly not unquestionably progressive.

We did all agree that this and other voice issues are very complex ones and it would need another arena to investigate them fully. It is the fact of the disagreement around voice, rather than its exact nature, which is relevant to questions of assessment raised in this study.

---

[7] The external examiner subsequently agreed that not all ethnographic research report writing should be categorically 'warm' or 'advocatory'. He objected to our use of the term 'hot' because 'research is not about sensationalism!' He added,

> The representation preference of an author is his or her stylistic choice. But if the very representation style excludes/marginalises readers (and particular groups of readers) then it needs to be questioned (see Samuel, 1988). Who is postgraduate research for? The existing academic world? By being warm or inviting, one's circle of audience is likely to be expanded.
>
> (Samuel, personal communication, September , 2002)

Implications for assessment practice

In summary, our research has produced the following findings: in investigating the evidence of inter-marker [un]reliability, we discovered shared categories, differences in how these are weighted and two unresolved issues. As a result, we conclude that the 'consistency' of assessment within and across universities aspired to by quality assurers (such as the HEQC in the South African case) is difficult to achieve.

To address the problem of [un]reliability which we believe has equity implications for students, we developed a set of banded criteria (Table 2) and agreed, provisionally, to weightings for the different categories of assessment. Further investigation will establish whether this instrument does in fact produce greater inter-marker reliability. Use of the instrument by different markers will also increase our understanding of whether or not the criteria have been adequately specified. In the meantime, this specification makes our assessment criteria transparent for students, staff and the external examiner.

The use of language and the use of voice remain unresolved issues. These have been addressed in different ways. With regard to language, we were able to agree on the end points of a language continuum – what constitutes a distinctive use of language, what is passable and what constitutes a failure. We have a shared understanding that there is a continuum between these end points and that it is impossible to provide fixed, graded points along this continuum because of the difficulties of assessing language in isolation from the other categories of assessment.

The question of voice proved to be more intractable because of the fundamental philosophical disagreements. While we have provided the poles of the cline, there are some staff who do not position themselves at these poles nor do they imagine themselves at some position on the cline. There is also no way of including voice in a set of banded criteria, as the poles of the cline are not tied to grades in the way the language poles are. At this stage multiple marking and the averaging of the marks awarded across markers seems to be the fairest way of dealing with differences across markers. Further research into the use of voice, needs to focus on students' writing as the data to be investigated.

What we have uncovered, in retrospect, may seem completely obvious, but until we undertook this investigation through the different forms of analysis, we were unable to articulate it. The process of making our implicit assumptions explicit has led to a greater awareness of the factors that produce and may continue to produce inter-marker [un]reliability in assessing students' research reports.

References

Biggs, J. 1999. *Teaching for quality learning at University*. Buckingham: Society for Research in Higher Education and Open University Press.

Brown, S, Race, P & Smith, B. 1996. *500 tips on assessment*. London: Kogan Page.

Brown, G with Bull, J & Pendlebury, M. 1997. *Assessing student learning in higher education*. London: Routledge.

Clifford, J & Marcus, G. 1986. *Writing culture: The poetics and politics of ethnography*. Berkeley: University of California Press.

Collis, KF & Biggs, JB. 1983. Matriculation and requirements and cognitive demands in universities and CAE's. *Australian Journal of Education*, **27**,41-51.

Freeman, D. 1998. *Doing teacher research: From inquiry to understanding.* Canada: Heinle and Heinle.

Geertz, C. 1973. *The interpretation of cultures*. New York: Basic Books.

Gipps, C. 1994. *Beyond testing*. London: The Falmer Press.

Maton, K. 1998. *Recovering pedagogic discourse: Basil Bernstein and the rise of taught academic subjects in higher education*. Paper presented at the 'Knowledge, Identity and Pedagogy' Conference, University of Southhampton, March.

Moore, R & Muller, J. 1999. The discourse of 'voice' and the problem of knowledge and identity in the sociology of education', *British Journal of Sociology of Education*, **20**(2), 27-34.

Smout, M. (ed.). 2002. *Quality assurance in South African Universities*. Pretoria: SAUVCA.

University of Western Australia. 1999. *Issues of teaching and learning*. Available url: (http://www.csd.uwa.edu.au/newsletter/issue0899). Accessed 16 April 2002.

Walkerdine, V. 1997. *Daddy's girl: Young girls and popular culture*. Cambridge: Harvard University Press.