Validating academic assessment: a hermeneutical perspective [1]

Johann Graaff, University of Cape Town

graaff@humanities.uct.ac.za

Yvonne Reed, University of the Witwatersrand

yvonne.reed@wits.ac.za.

Suellen Shay, University of Cape Town

sshay@ched.uct.ac.za

Abstract

This article addresses the nature of validation in **assessment**, that is, the questionof what we know, and the processes by which we come to know, in assessing student work. Interest in this question started with a panel discussion at the Kenton Education Conference between the three authors of this article. This article is a continuation of that discussion. It begins by drawing on the basic distinction between the hermeneutics of faith and the hermeneutics of suspicion, first set out by Paul Ricoeur. These are two ontological moments in social science theory, in everyday life, in teaching and in **assessment**. They cannot be separated. Nevertheless, and quite problematically, much of the **assessment** literature, and much **assessment** activity, ignores the first and emphasizes the second. In addition, there is significant **assessment** activity in teaching which incorporates implicitly and silently, non-cognitive and situational factors, based on the hermeneutics of faith. Our question is: How is one, then, to validate judgements made in this post-positivist mode? How is one to assess the assessor? We conclude the paper with tentative suggestions of how criteria drawn from qualitative research and from psychotherapy can be helpful in this regard.

Introduction

At the Kenton Education Conference in October 2003, the authors gave a panel presentation entitled, "**Assessment** in Higher Education: A matter of 'taste'?" Reed and Shay, drawing from previously published research (Reed *et al.*, 2003; Shay, in press), each presented a case study of an **academic** community of practice struggling to achieve consensus in relation to their interpretations of student performance. A common theme which arose from the case study presentations is that the **assessment** of complex tasks, for example, honours projects and theses, is an interpretive process which draws on a variety of sources of evidence. Even where marking criteria are made explicit, they cannot do justice to the complexity of the

interpretive frameworks which assessors are implicitly drawing on. In his contribution to the panel, Graaff applied a critical **hermeneutical perspective** to illumine and explore in more depth some of the complexities of these interpretive processes. This article captures (and extends) this conversation.

Both the Reed *et al.* and Shay articles address the issue of inter-marker reliability (IMR), (that is, the degree to which assessors agree or differ in their **assessment** of students' performance) but they do so in quite different ways. Shay's account takes a sociological approach, exploring why academics agree and why they disagree in their marking. In strong post-positivist mode, she argues that the time for objective, absolute **assessment** criteria, and hence for complete consensus, has passed. But this does not mean that anything goes. She ends on an optimistic note, following Bent Flyvbjerg (Flyvbjerg, 2001), by showing the possibilities of a new kind of dialogue between academics which is collaborative, contextual, flexible, value-sensitive and community-forming; abandoning the old certainties of positivism, but without lapsing into normless relativism.

The Reed *et al.* article, *[Un]reliable **assessment**: A case study*, by contrast, takes a more ethnographic approach. The authors present a discussion of an investigation undertaken by a group of colleagues in one **academic** department into certain anomalies which arose in the **assessment** of a particular group of post graduate students' research reports. They 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 impeding inter-marker reliability. Through a content and discourse analysis of the different assessors' written reports, they attempted to uncover the implicit **assessment** categories and criteria which assessors were working with in their assessments. They discovered shared categories and criteria as well as differences in how these criteria were weighted. In the interests of equity and increased inter-marker reliability they developed a set of banded criteria on generic features of research reports. The article also presents a discussion of two issues that the group did not resolve: the weighting to be given to students' use of language, and the role of the writer's 'voice' in the research report. Their conclusion is more pessimistic, that consensus between markers is unlikely. Even where they do agree on appropriate criteria for **assessment**, which they do not always do, they may not agree on the ranking and priority of these criteria, nor on how to define the criteria.

Four main points emerge from the Reed *et al.* and Shay research. First, they both focus on events which are remarkable in their rarity in academia, namely, detailed conversations between academics on marking criteria. Not to put too fine a point on it, academics hardly ever engage with each other on the detail of how to mark student work. Even in situations where such conversations are institutionalized, like those with external examiners, or between co-markers, they are often superficial and brief. The reasons for this are fairly evident. Such conversations do not enjoy high priority outside of the group of professional educationists (nor do they feature in performance appraisal checklists), and they are immensely time-consuming. As Shay shows, most young academics are socialized into the area of marking criteria by just doing it. There are precious few explicit guidelines, workshops, or mentoring opportunities on marking. What Reed *et al.* do in their article is break down the criterion of

'voice' into four different possible interpretations. This could be replicated for any number of other conventionally unexamined criteria like 'structure' or 'relevance'. In short, the possibilities for disagreement, and then exploratory dialogue, are infinite.

Second, inter-marker dialogue can as easily lead to exacerbated conflict, hardened discord and fragmentation of community as to consensus or agreement. (Shay notes the interesting possibility of agreeing on a particular position for divergent reasons – which she calls agreement rather than consensus.) In this regard, Flyvbjerg distinguishes the varying approaches to dialogue between Foucault and Habermas. Where Foucault starts from the assumption that dialogue cannot but involve conflict, negotiation and struggles over power, Habermas believes that full consensus is something that is desirable and achievable (Flyvbjerg, 2001, chap 7).

Third, and important for the later development of the argument, both Reed *et al.* and Shay note the particular position occupied by the research supervisor as assessor. Supervisors typically get themselves involved in the personal details and activities of students/dissertation-writers. Supervisors' take on the process of **assessment**, in consequence, is often then significantly coloured by these different circumstances. This has important implications for **assessment**, as we shall see.

Fourthly, while both the Reed *et al.* and Shay arguments point to the complexity of assessors' interpretive frameworks, their case studies shed little light on the full range of criteria which assessors are drawing on. Assessors in their case studies are busy debating cognitive criteria, such as the 'value and priority of structure', or 'voice', or 'conceptual sophistication'. We argue here that a great deal of **assessment** happens, silent and unacknowledged, around what we call, for want of a better term, non-cognitive or contextual criteria. That means those considerations of student process, of application, stamina, courage in the face of adversity, accidents and tragedies. It is important that we investigate ways to think about these and their role in **assessment**. And if cognitive **assessment** criteria are seldom reflected on, it is almost certain that the non-cognitive will be completely invisible.

Hermeneutics of suspicion and hermeneutics of faith

At this point let us step back and sketch a broader frame within which to locate the argument that follows. Paul Ricoeur (1981) makes the fundamental distinction in the social sciences between the *hermeneutics of suspicion* and the *hermeneutics of faith*, or put differently, the distinction between a critical, judgemental approach, and a sensitive, empathic one. Quite a lot of the **assessment** literature, even when it emphasises the importance of making **assessment** criteria explicit to learners, operates from a critical judgemental approach (e.g. Brown and Knight, 1994; Knight, 1995). This is, after all, what assessors do – they judge, using a set of commonly accepted criteria. But there is another side to **assessment** which is empathic and communicative, and that entails a different set of criteria which are seldom spelt out.

In an early article Ricoeur points to a fundamental fissure within the social sciences. In theoretical terms this fissure can be simply indicated as the split between critical theory and hermeneutics, and to be historically concrete, the debate between Habermas and Gadamer. For Ricoeur this was part of a much bigger debate between what he termed the *hermeneutics of faith* as represented by phenomenology and Gadamer's hermeneutics, on the one hand, and the *hermeneutics of suspicion*, as represented by Marx, Freud and the Frankfurt school of critical theory, on the other (Ricoeur, 1981). The central difference between these two, said

Ricoeur, lay in their approach to texts. It is important that they are both, for Ricoeur, forms of hermeneutics, and therefore distance themselves from positivism. Despite its earlier positivist tendencies, critical theory, in its Habermasian form, rejects the subject-object split, and accepts the unavoidability of interpretation in communication.

The hermeneutics of suspicion takes its cue from Marx's notion of false consciousness. It starts from the principle that certain texts are biased through their ideological function of disguising social inequalities. Thus, for example, under capitalism the ruling class is concerned to justify its position of power in society and to perpetuate that position. In this sense what people say about themselves and about the world cannot be trusted. It is, in Habermas's terminology, 'systematically distorted'. It is the role of the social scientist to debunk this language and to show the reality of exploitation and oppression lying behind it.

The hermeneutics of faith, on the other hand, starts from the assumption that people's language is difficult to understand. This language has multiple levels and meanings which must be carefully and respectfully unravelled, that is, it must be taken on their own terms. This approach is characterized by "an attitude of care and concern. . . so as to allow the full weight of the (sacred) message to appear" (quoted in How, 1995, p.18) It is only with such a sympathetic attitude that the full meaning of language *can* be uncovered.

While the Habermas-Gadamer debate is here presented as something which happens out there in the world of theory, the two modes of interpretation being discussed are also embedded in everyday modes of communication. They are both constitutive of social reality. One is a judging, intervening, harsher, robust activity. The other is an empathic, non-judgemental, respectful, caring activity.

Anthony Giddens makes a similar distinction between mutual knowledge and common sense in the conduct of social science (Giddens, 1984). For Giddens, these are two perspectives on social reality which researchers conventionally take. Mutual knowledge indicates the way in which researchers must share the world of social actors in order for social science to be possible. It is a condition for any kind of engagement with the world of social actors.

---

[1] Historically, the debate occurred in the late 1960s and early 1970s. Our argument here revolves around an article originally written by Paul Ricoeur in 1973 (and later translated into English by John Thompson in 1981 (Ricoeur, 1981)). Ricoeur's article was written as a comment on the debate between Hans-Georg Gadamer and Jurgen Habermas which stretched over the four years between 1967 and 1971. Gadamer's seminal work, *Truth and Method*, appeared in German in 1960.

Common sense, by contrast, indicates a distancing, a critical stance, between observers and social actors. Researchers regard everyday knowledge as common sense when they contrast it with social science. Common sense, then, for Giddens, comprises "the prepositional beliefs implicated in the conduct of day-to-day activities" (1984, p. 336). Giddens's mutual knowledge/common sense distinction exactly parallels Ricoeur's hermeneutics of faith/hermeneutics of suspicion distinction – with one important difference. For Giddens the critical moment which social science necessarily implies is wider than just ideology critique. Habermas, and the neo-Marxist current within which he operates, focus strongly on the unveiling of the hidden places of power in society. To that extent their **perspective** is a narrow one. It excludes important (hidden) aspects of social action which are not necessarily related to power. For this reason we prefer Giddens's notion of sociology's critical stance. Now, there is a very complex relationship between these two modes, the empathic and the critical. From one **perspective** these modes are mutually exclusive. The subtleties of someone's psyche and meaning cannot be explored if there are preset categories of power, class or gender which intervene to foreclose further exploration. (This dilemma is mirrored in the Catholic priest's rule that one is forbidden from revealing crimes that have come to light during confession. Or the psychotherapist's ethic of confidentiality.) You either listen or you judge. You cannot do both simultaneously. Conversely, people cannot act within society in a moral sense without making some kind of judgement that is, drawing back from the empathic mode.

From another **perspective**, these modes often operate in tandem. Teaching, for example, is an activity where one individual is elevated in status above others, as an authority, who makes critical judgements on (i.e. assesses) the progress and worth of other individuals. Even in a gentler child-centred light, teaching is also comment, commentary, information about other opinions. At the same time, teaching is also a form of communication where delicate and sensitive listening is needed. It would be difficult to have the one without the other (although obviously they occur in different mixtures in particular individuals and in particular acts). At a very elementary level of operation, also, it is evident that one cannot discern the exercise of power in social relationships without a prior interpretive act. (How would one know that a particular act was one that entailed power?) In many instances, then, the critical and the empathic modes are moments of the same act. They cannot be separated.

For Giddens, likewise, as indicated above, there is a wider critical **perspective** (than power) implied in all of social science. It is extremely difficult not to participate in some kind of comment or commentary on people's every day acts even in the most descriptive and ethnographic of research. Simply to translate everyday language into sociological terminology is already a transformation of it, offering a comment on it. To refer to people, for example, who call themselves 'freedom fighters' as 'members of a political movement' already changes the emotional colour of the terminology (Giddens, 1984, p.337).

There are extremely interesting conundrums thrown up by this juxtaposition of perspectives. Critical theory (of the Habermasian or neo-Marxist variety), which is nothing if it is not critical of unequal power, cannot operate without setting someone up as judge who has the

(unequal) authority to be critical. Put differently, critical theory cannot be critical of inequality in society without an acceptance of inequality at its own core. Conversely, hermeneutics, which, famously, enrages critical theorists by defending the value of authority in society, insists that interpretation cannot be properly done without a supreme act of humility.

These juxtapositions, parallels and conundrums translate directly into the sphere of **assessment**. As intimated above, summative **assessment** (e.g. for selection and certification) necessarily has a judgemental, differentiating and critical aspect to it. By its practical operation, it lives out the need to discriminate between individual students, categorizing them sometimes on quite crude and predetermined criteria into groups of 'better' and 'worse', 'pass' or 'fail', 'second class' or 'upper second'. **Assessment** in this mode is also then an active intervention in the world, an attempt to make things better, predicated on a moral judgement of what is more valued or less so.

Formative and diagnostic **assessment**, by contrast, tends in the other sensitive, listening and empathic direction with judgement and critique muted, although not suspended. It is an attempt to make things better, but, more importantly, an attempt to enter into the realm of the other individual, to share those particular values and assumptions. In Rogerian psychology, it is the 'unconditional positive regard' which lies at the heart of good therapy and good communication (Webb, 1996). Ideally, there are no preset criteria here, but multiple and various criteria which flow from each particular situation on the way to individual self-knowledge. In this empathic mode, the word **assessment** is itself inappropriate.

But, much like the world of social interaction, it is very difficult to separate these two moments of **assessment**. One needs to listen in order to be able to judge. The assessor needs sufficient sensitivity to enter the world of the student text before any judgement can be made, to know at all what learners' texts are saying. (And it is a matter of some importance whether one is listening only for the markers of preset cognitive criteria, or whether one is open to other signs too, as we shall see in a moment.) It is also a central aim of **assessment** not just to make a judgement but also to make a difference, to communicate that judgement in a way that produces change. Assessors would be bad teachers if they used language which students could not connect with.

Conversely, one cannot communicate in **assessment** without making a judgement. That would seem obvious and even tautological. After all, **assessment** is by definition a kind of judgement. The difference between formative and summative **assessment** is not that one of them is not judging. They are simply different kinds of judgement. Thus two modes – empathy and critique, listening and intervention, interpretation and judgement – clash and interweave and complement each other. They do this in research, in teaching and in **assessment**.

We now turn to the issue of the criteria which academics draw on in their **assessment** of student performance, an issue explored by both Reed *et al.* and Shay. The following section examines a case study which highlights the influence of non-cognitive or contextual factors

on assessors' judgements. More intriguing are the criteria by which one considers the validity of this kind of judgement. Maureen Angen and Paul Ricoeur have proposed criteria for **validating** just this kind of qualitative interpretation. We discuss these below.

Assessment as Communication

Let us consider then a fictitious but very recognizable concrete case.

> Student N's essays normally score in the region of 65-70%. However, now and again, she hands her assignment in very late, earning a penalty of 50% of her mark. At the end of the semester, her average mark is 66.5%. Taking the penalties into account, however, her course mark drops to 42%. She regularly arrives some 20 minutes late for class. In the exam she again arrives 20 minutes late. She fails the exam. If her coursework mark is taken as 66.5% she would qualify for a re-exam and have the possibility of passing the course. She has failed both of the other BSocSci courses she is following. We call her in. She explains that she did not want to come to university. Her parents forced her. However, she has now turned over a new leaf and has decided to do her best. She wants very much to rewrite the exam and pass the course. She mentions that in the interim she has also passed a supplementary exam in another subject. She means to signal to us that she is on her new path.

The point to make about this case is not its complexity or its unusual nature. As assessors we are required to make judgements of this kind all the time. In exam committee meetings to allocate final marks, assessors quite frequently say things like: "This is a truly lazy student. She does not deserve to pass", or "She went through a really difficult patch when her mother died. She needs to be treated sympathetically." Where students hover on the edge of categories, pass or fail, second or third class, these considerations quite frequently tip the balance. They are an integral but invisible part of the **assessment** process, invisible because there are no 'standard' criteria for making this kind of judgement, and because no one does research about this aspect of the process.

And, yes, the **assessment** is complex because it entails delicate, albeit unthinking, moral judgement, entangled in the specifics of context and the individual case. It is then entirely to be expected that thesis/dissertation supervisors find themselves swayed by precisely these factors when assessing the work of their own students. They are, over a period of time, immersed in the detailed colours and textures of personal student endeavour, of patience and stamina, of enthusiasm or lassitude, of obstructing or enabling events in everyday lives. Clearly their judgement of student work comes from a very different angle from that of an impersonal external examiner.

And how should we think about the judgements of supervisors then? Should they be taken as better or worse judges of student work? Should supervisors be set aside as assessors, or given the deciding vote where there is lack of consensus between markers?

Just as universities (and within them, faculties, schools, etc.) differ in terms of who assesses post-graduate research reports, dissertations and theses and in terms of the criteria used in this **assessment**, so too do they differ in terms of how the outcomes of the **assessment** are reported to students. In some instances students receive a copy of a detailed **assessment** report and in others, only a mark or a mark and a brief summarising comment, with the more detailed report of both internal and external examiners being filed in a department, school or faculty archive. If, in the interests of equity, fairness and accountability full reports should be made available to students, then perhaps the collaborative conversation among academics advocated by Shay (2003) is a prerequisite for the development of both shared understandings about **assessment** and shared discourse(s) for use in writing such reports.

The inclusion of non-cognitive factors in teachers' **assessment** of students' work connects to one of the central concerns of the Reed *et al.* article (2003). This is a concern about **assessment** equity: how to be fair to each student, which may involve 'treating' students unequally in terms of providing extra scaffolding or overlooking linguistic errors in the writing of ESL/EFL students. The suggestion, that a teacher-assessor be immersed in the detail of his/her students' lives, may help to address the critical question of "whether the assessors' interpretations of the criteria or standards are culturally or socially discriminatory or whether they are inclusive of the clientele" (Freiberg, 2001, p.290). However, also it raises a number of other questions which include the following:

1) How feasible is such immersion in the life of each student if a teacher is working with large numbers of students (as is common in tertiary level undergraduate classes)?
2) Do all students wish to/find it possible to offer such personal detail to their teachers?
3) Can teachers always evaluate accurately the 'truth' of such personal detail (back to the hermeneutics of suspicion)?
4) How is knowledge of the person to be 'treated' equitably across students?
5) How is this knowledge to be represented in an **assessment** report (i.e. are changes/additions to **assessment** discourse(s) required)?

In our case-study, the student 'opens up' to her lecturers and they make a positive response which is subsequently vindicated by the student's improved work. But what of the students who do not let their teachers into their lives for whatever reason? As an example, one of us was concerned about the marked deterioration in performance on assignments of a student in a distance learning programme. Despite an invitation to her to discuss any difficulties she was experiencing, she remained silent. At a subsequent residential component of the programme, it transpired that this student had a son whom she was nursing through the terminal stages of HIV-AIDS yet she asked for no special consideration in circumstances where such consideration would have been justified. Another student in the same in-service teacher education programme claimed that it was not possible for her to complete some of the assignments because conditions at her school were too difficult (in terms of limited resources and lack of co-operative colleagues). One might have accepted this claim and attempted to modify the assignments or assess her work more 'leniently' than that of other students.

However, there had been another teacher from the same school in the programme in the previous year who had painted quite a different picture of the teaching and learning infrastructure. Here it was possible to challenge the claim for special treatment.

It is right that immersion in the life of the student is more likely to happen (though not inevitably) in the interactions between supervisor-supervisee during the course of a post-graduate research project than in the teaching of large classes. In some universities, both in South Africa and elsewhere, it is policy for the supervisor to also be the internal marker while in others the internal marker is an **academic** who has had no involvement in the student's research. Here again the issue of equity and fairness remains unresolved or perhaps becomes a burden for the external examiner to shoulder. It is unresolved because different supervisors will 'listen' to their students in different ways and give different value to such 'non-cognitive' variables as 'persistence', 'improvement' and 'dedication'. From a **hermeneutical perspective** these practical **assessment** dilemmas, for example different interpretations based on different values, are inevitable. A **hermeneutical** approach to **assessment** would suggest that because of different perspectives two different assessors will not see the same thing. Different perspectives are in part shaped by different purposes. While it is common to refer to the different purposes or functions of **assessment**, for example, summative and formative, the case study and illustrations above point to multiple and more subtle functions which **assessment** serves. On the one hand, a mark is a measure of product against some tacitly agreed notions which an **academic** community (within the wider context of the disciplinary field) holds about 'good' research. On the other hand, the mark is a measure of process, a communicative exchange between an assessor and a student. These functions create both intra-marker and inter-marker tensions because they privilege different contextual considerations, different forms of evidence. In its function as a measure of product, **assessment** interpretations may foreground products (e.g. the research project) and background processes (e.g. the drafts, the class discussions, one-on-one time with the student). In its function as a measure of process, **assessment** may privilege evidence from the student (e.g. their effort, progress, conditions external to the classroom). Supervisors feel the tension as these different functions pull on their judgements. In privileging one function over the other, supervisors and external examiners may come to different interpretations.

These functions – the **assessment** of process and product – and the kinds of evidence which they privilege should ideally be held together in order to strengthen the soundness of our interpretations of student performance. It is thus not a case of whose **perspective** is most valid – the external or the internal examiners' – but how these perspectives can be held together to strengthen validity.

Validating hermeneutic activity

In the remaining sections of this article, we consider two possibilities for **validating** such contextual aspects of **assessment**, one from the area of qualitative research and one from psychology. We go outside the sphere of pedagogy for the reason that there is very little in

the way of pedagogical literature. We start with a piece by Maureen Angen (2000) who considers the difficulties of **validating** qualitative research. It helps us because qualitative research is likewise in the business of non-positivist validation. Angen breaks away from notions of independently verifiable reality, and begins to think very creatively about how validation operates in this heretofore grey area. We continue this investigation by considering another area which is analogous to **assessment**, namely psychotherapy. Here we consider Paul Ricoeur's discussion of validation in Freudian psychology, and the issue of how to validate therapeutic activity, with the aim of applying that to **assessment**.

Let us start then with Maureen Angen's discussion on **validating** qualitative research. This discussion is difficult because validation is a term that derives from positivism. It implies a confirmation, a checking, that the results of research coordinate with an independent, objective, constant reality. Hermeneutics, however, denies the possibility of separating subject and object in this way. More importantly, it emphasizes the fact that subject and object influence each other. Two different observers are not observing the same thing because, as a direct consequence of their different presences, they are looking at different things. There can, therefore, be no checking for the independent existence of this relationship. Angen shows very clearly, however, how much qualitative research still clings to 'residual realism' or 'subtle realism' via such notions as triangulation, clarification of researcher bias, external audits and peer review. If one then rejects the positivist basis for validation on these grounds, the question is: what other grounds are available for 'checking', and is checking even the right term?

Angen argues that interpretive theory recognizes that social reality is intersubjectively negotiated, shifting, morally implicated, prejudiced. On this basis, validation can indeed be judged by:

a. **ethical aspects**: "Interpretive research should provide a thoughtful, caring and responsible answer to the question, 'How do we become more fully who we are?' as human beings" (Angen, 2000, p.388). Hence the research needs to be *beneficial* to people, *useful* – "a piece of research unfolds into the future as the interpretation is taken up by the community of practitioners" (*ibid*.), it must be *generative* of new ideas and questions (what Angen calls it 'rhizomatic' and 'voluptuous' validity), *transformative* of our actions;
b. **substantive aspects**: writing must 'do justice' to the thickness of the situation, self-critical reflection must show how understandings have shifted, it must be trustworthy, give evidence of a feeling of authenticity, an 'aha' experience. Here Angen uses the term 'spontaneous validity';
c. **qualities of the researcher**: researchers must be resilient, patient, persistent, meticulous, passionate, and personally involved.

Now, this is a considerable shift from older notions whereby validity was checked against independent reality. Here we are dealing far more with issues of ethics and accountability. The question to be asked of research here is not: "Does it accord with reality?" but rather,

"Does the researcher comport him/herself in a responsible, accountable manner? Does he/she do good or does he/she do damage?"
How helpful are Angen's criteria, then, for **assessment**, that is, **assessment** in a hermeneutic sense? Well, very helpful indeed. Let us consider each of Angen's three criteria. If what comes out of classrooms is a product of the interaction between very particular teachers with very particular students rather than an objective and independently measurable reality, then, firstly, it is critical that **assessment** be done in a beneficent rather than combative, hostile or destructive mode. If **assessment** is careful listening to the multiple chords of individual complexity, then, secondly, it is critical that teachers have an intimate and 'thick' knowledge of students' lives. If, thirdly, teachers are to act according to these first two criteria, then it is strongly implied that they be also 'resilient, patient, persistent, meticulous, passionate, and personally involved'.

The interesting question that arises from all of this is how one assesses the assessor. How is one to decide (validate) whether a particular teacher is benevolent and transformative, immersed in the detail of his/her students' lives, and acting in a patient and meticulous way? Should teachers be required to write biographies of individual students? In qualitative research, presumably many of these attributes are discernible from the texts that researchers produce. The texts that teachers produce, their teaching performances, on the other hand, are not so easy to pin down and read off the page. Nor are the requirements for public accountability in these aspects particularly stringent.

Assessment as psychotherapy

If qualitative research is one exemplar for **assessment**, let us now turn to another, that of psychotherapy. Within therapeutic practice there is, depending on the theoretical paradigm within which one works, a corpus of criteria and guidelines for what constitutes good interpretation, and for validity checks as well. Within Rogerian psychology, for example, one of the standard formulae for communication is the process of 'reflecting back' what a client is saying. Typically a therapist will respond to a client with, "In other words, what you are saying is. . ." It is a measure of the quality of his/her interpretation that a client is able to say in turn, "Yes, that's right." More penetratingly, a therapist might also say, "In other words, what you are *doing* is. . ." Ideally a client would then say, "Gosh, I never thought about it like that!" Habermas mentions a similar check on therapeutic interpretation, namely, that the client can recognize and respond to the interpretation, that it 'rings a bell'.

Ricoeur, in his examination of proof in Freudian psychoanalysis, has a more comprehensive take. For him, there are four criteria against which to judge the quality of therapeutic intervention: (a) that it conform to Freudian theory; (b) that it interpret according to the (theoretically determined) rules of interpretation, as in dream analysis; (c) that it is effective in bringing about change; and (d) that the interpretation constitutes a narrative which makes sense (Ricoeur, 1981). The Rogerian interchange set out above would then conform to

criteria (b) and (c). It interprets according to certain rules, and it is effective in making an intervention.

Transposed into the **assessment** situation, one can then also ask three questions:

(a) Is the teacher/assessor listening in a way that conforms to accepted standards? Is it good or bad listening?

(b) Does the intervention make a difference?

(c) Can the teacher/assessor's act be explained to someone else in a way that makes sense?

In class discussion, then, a teacher might say in response to a student, exactly as a therapist does, "What you are saying is . . .?", and furthermore, "What you are *doing* is …?" It would be an important measure of that interpretation, that a student is able to say, "Yes, that's right", or more valuably, "Yes, that's right!! And what's more is . . ." Likewise it would be important that a student also said, "Gosh, I never thought about it like that before", and then go on to adjust his/her practice (in writing essays or arguing a point).

Let us now put these two sets of criteria together, Angen's and Ricoeur's. For our purposes, they can be collapsed into four main questions which can be asked in order to validate and interrogate interventions in teaching and **assessment**:

a. Is the aim of the intervention to do good or to do damage? Is it constructive or simply punitive?
b. Is the intervention based on 'thick' knowledge of the student?
c. What is the impact of the intervention?
d. Is the intervention discussed with other teachers or in a public forum? Does it make sense to someone other than the intervenor?

If we accept the **assessment** of complex tasks as a socially situated, interpretive act, the implication is that this requires a careful re-think of the basis on which **academic** communities are confident about the interpretations they make of their students' performance – interpretations which have serious implications for students' lives. The questions above and particularly the last one – whether the intervention 'makes sense' to someone else – point to a crucial **validating** criterion. The criterion is the existence of a **validating** community – as Richard Bernstein (1976, p.111) puts it, "a community of inquirers who are able, willing and committed to engage in the argumentation" . What is required is a community of assessors who are committed to dialogue where there is "choice, deliberation, interpretation, judicious weighing and application of 'universal criteria', and even rational disagreement about which criteria are relevant and most important" (Bernstein 1983, p.172).

Are academics and their institutions willing to engage in this kind of dialogue in the face of so many competing priorities? Some will, most will not. Under these circumstances higher education has no choice but to remain extremely self-critical about its classificatory systems and the consequences of these classifications for students and the broader society.

Conclusion

From problems of inter-marker reliability leading to dialogues around cognitive criteria, to problems of judgement on non-cognitive factors influencing **academic** matters. This is very evidently an uncharted area. Trying to map guidelines from qualitative research and from psychotherapy on to issues of **assessment** is a sign of just how uncharted. For validation it means moving from the old and secure practice of checking on an independent external reality to the more nebulous one of reflecting on ethical propriety. In ancient times mariners stayed away from such unmapped areas fearing the presence of dragons. The chances are, quite a few academics would shy away from this area for equally powerful reasons, that it contains far too many grey areas, opportunities for conflict, even dangers of litigation from students – there is nothing more academically dragon-like than that! In the end, mariners learned that they would not fall off the edge of the world, and kept going, right round the world. Are academics that open to new thinking? Furthermore, are institutions open to recognising the time that academics need for dialogue with colleagues, for formative feedback to students and for the preparation of summative reports which are both 'empathic' and 'critical' in relation to the work submitted for **assessment**?

Assessment is a knowledge activity that is central to the **academic** enterprise. The issue of **validating assessment**, that is, the way in which we test our judgements about students, very evidently then goes to the heart of the theme of the 2003 Kenton Conference: what do we know? and how do we know it? This article is part of an ongoing conversation which explores the production and the critique of knowledge in a world that has gone beyond positivism.

References

Angen, M. 2000. Pearls, pith and provocation: Evaluating interpretive inquiry:

Reviewing the validity debate and opening the dialogue. *Qualitative Health*

Research, 10: pp.378-395.

Bernstein, R. 1983. *Beyond objectivism and relativism: Science, hermeneutics,*

and praxis. Philadelphia: University of Pennsylvania.

Bernstein, R. 1976. *The restructuring of social and political theory*. New York:

Harcourt Brace Jovanovich.

Brown, S. and Knight, P. 1994. Assessing learners in higher education. Kogan

Page: London.

Flyvbjerg, B. 2001. *Making social science matter: Why social inquiry fails and how it can succeed again.* Cambridge University Press: Cambridge.

Freiberg, J. 2001. Criteria-based **assessment** in senior high school English: Transcending the textual in search of the magical. In Freebody, P., Muspratt, S. and Dwyer, B. (Eds). *Difference, silence and textual practice*. Cresskill (NJ): Hampton Press.

Giddens, A. 1984. *The constitution of society.* Cambridge: Polity Press.

How, A. 1995. *The Habermas-Gadamer debate and the nature of the social: back to bedrock.* Aldershot, Hants: Avebury.

Knight, P. (Ed.) 1995. **Assessment** for learning in higher education. London: Kogan Page.

Reed, Y., Granville, S., Janks, H., Makoe, P., Stein, P., Van Zyl, S.W. and Samuel, M. 2003. [Un]reliable **assessment**: a case study. *Perspectives in Education*, 21: pp.15-28.

Ricoeur, P. 1981. *Paul Ricoeur: Hermeneutics and the human sciences.* Cambridge: Cambridge University Press.

Shay, S. (in press). The **assessment** of complex performance: A socially-situated interpretive act. *Harvard Education Review*.

Ricoeur, P. 1981. *Paul Ricoeur: Hermeneutics and the human sciences.* Cambridge: Cambridge University Press.

Webb, G. 1996. *Understanding staff development.* Buckingham: Society for Research into Higher Education & Open University Press.