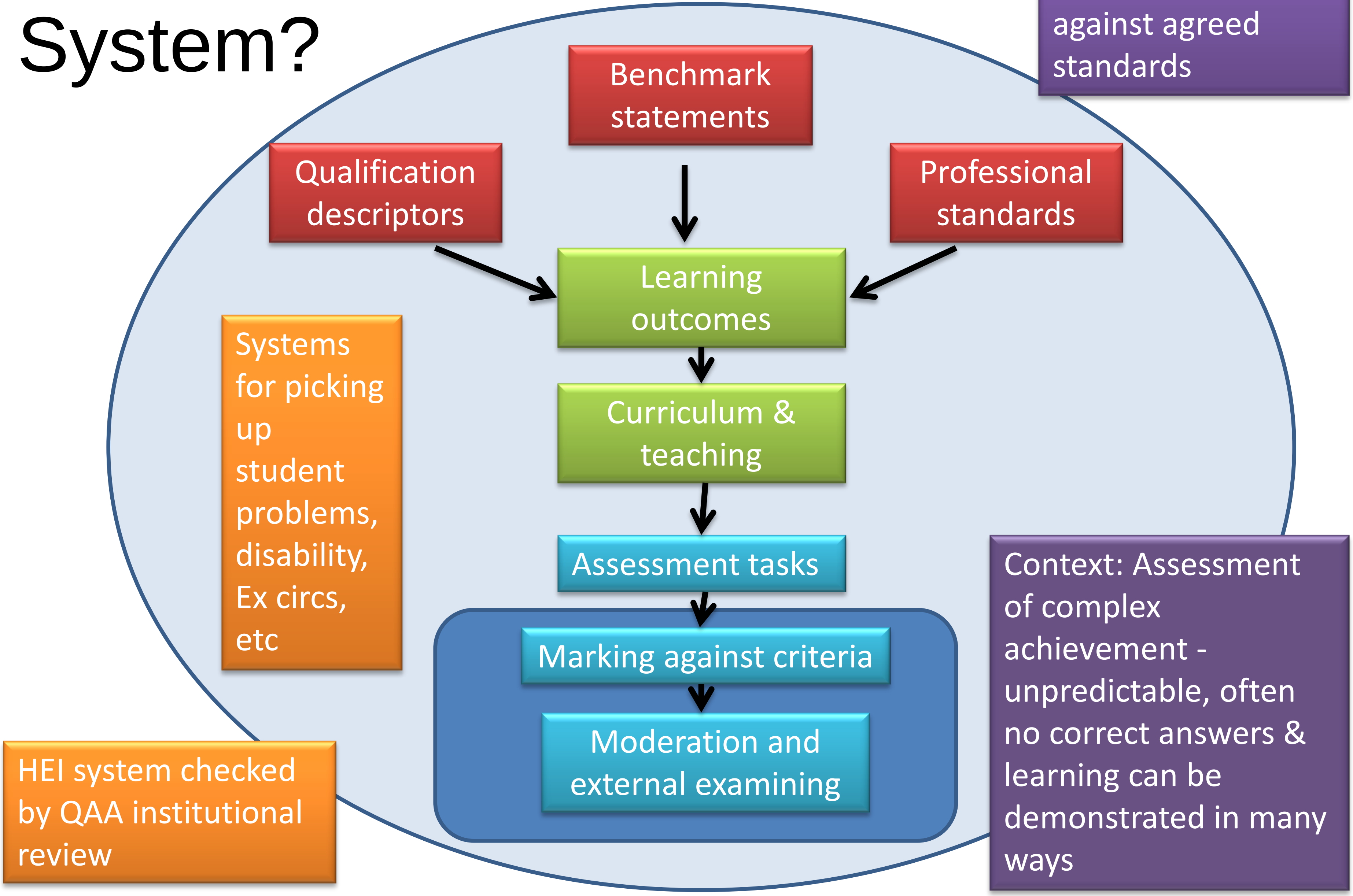
Investigating the mismatch between the policy and practice of assessment judgement in higher education

Sue Bloxham

UNIVERSITY of Cumbria

The Perfect System?



Result: consistent, reliable judgement against agreed standards

Benchmark statements

Qualification descriptors

Professional standards

Learning outcomes

Curriculum & teaching

Assessment tasks

Marking against criteria

Moderation and external examining

Systems for picking up student problems, disability, Ex circs, etc

HEI system checked by QAA institutional review

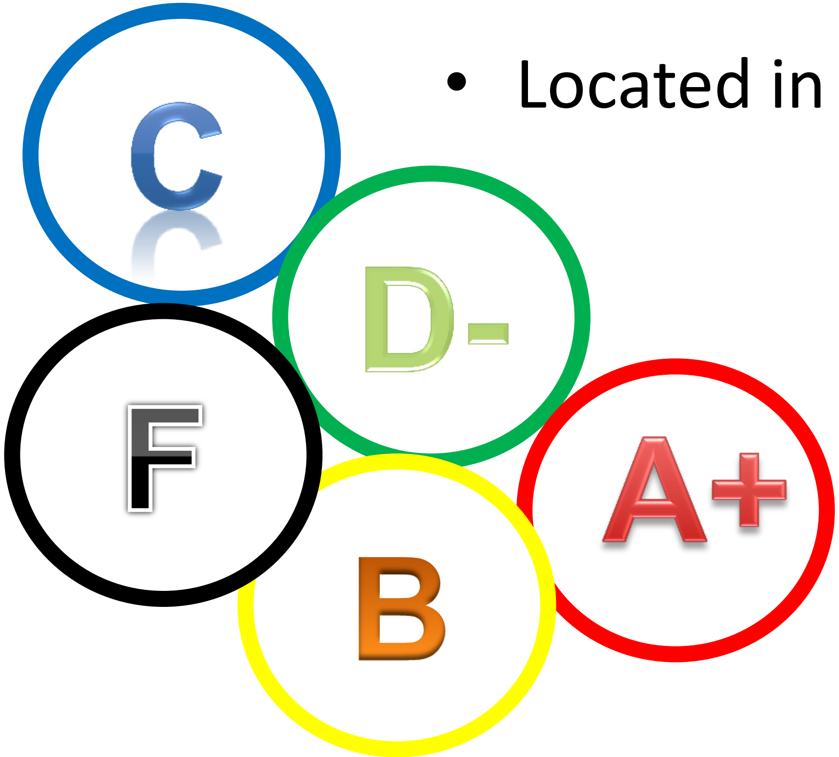
Context: Assessment of complex achievement - unpredictable, often no correct answers & learning can be demonstrated in many ways

Contention

- My contention today is that this perfect system has been developed with the admirable intentions of making assessment more transparent, fair, reliable and accountable and to maintain standards.
- However, the QA processes that we have adopted specifically for assessment are poorly matched to the nature of the learning assessed at this level and our knowledge of professional judgement in grading.

Techno-rational tradition in standards

- Perceives standards as something fixed, objective and measurable……. A GOLD standard. Based on assumptions that ‘knowledge is monolithic, static and universal’ (Delandshere 2001:127)



- Located in an objectivist epistemology

- Emphasis on **transparency** and creating **explicit** standards, e.g. professional standards, clarifying learning outcomes and assessment criteria

A broad alternative critique

This includes socio-cultural (Gipps, 1999), hermeneutic (Broad, 2003), social constructivist (Rust et al, 2005) and psychological perspectives (Brooks 2012).

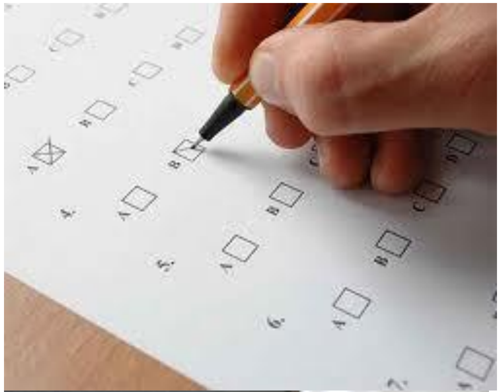
These approaches share an interpretivist approach to judgement and argue that the techno-rationalist' approach tends to ignore:

- beliefs, values, habits and purposes of tutors.
- the situated nature of grading decisions.
- the dynamic and contested nature of knowledge,
- the constructedness of knowledge.

There is both a theoretical critique and one based in empirical study of assessment and standards in action.

Main issues in research on use of standards in HE marking

- Meaning of standards socially situated and constituted
- Issues of complex judgement
- Lack of evidence of inter-subjectivity
- Unreliability well documented
- Lack of use of codified standards
- Heuristics and biases



Consensus through explicit standards?

- Difficulty codifying standards (Sadler 1987) – too general, abstract, hide complexity, mask diversity (Moss & Shultz 2001)
- Written statements need individual interpretation
- Holistic judgement – not using analytical standards in practice
- Prof. standards don't account for context
- Not grounded in empirical research

tda

CIPFA

NMC Nursing & Midwifery Council

QAA

Study 1: Data collection

- 25 tutors, responded to invitation;
- 3 Universities;
- Recorded marking two assignments, as they verbalise their thinking (A&D in 2s/3s);
- Followed by interview;
- Supplemented by researcher’s field notes;
- Subjects: teacher education; art & design, medicine, social science, humanities.

Surface characteristics

- Drives me crazy when students start sentences with numbers [] and consistent with all the other papers, all their references are put in the wrong place. (T23 medicine)

Holistic marking

Analysis supports theoretical arguments about the difficulty of analytical assessment of complex work:

Umm thinking about the essay for a while now and um glancing through it again, despite the comments that have been running through my mind about structure and the depth it does have you know, judging this from the point of view of a second year student rather than a usual history module it does have quite a lot of merit and I would not be disposed to give it a mark lower than a basic 2:1 but I would probably not go far above the 2:1 threshold. The essay has been fairly well researched I feel and although it deals in fairly general terms the sense I get is that it has used its research base fairly fully and certainly the research base stated in the bibliography is an enormous one. (T5)

Lack of explicit use of Criteria whilst marking

- This was rare and, when used, involved a ‘threshold’ rather than standards approach to criteria

Then she goes on to say why she chose the Vikings – because of its significance, its importance in understanding what it is to be British and where it fits into the standard Scheme of Work. All those are things in the criteria so again I’ll put a double tick in the margin just to remind me that I’ve ticked them off the criteria in my head as I do it. (T3)

Checking grades

Many tutors use explicit criteria/ objectives to check or confirm grades/pass as a final step:

OK. Now I step back from the essay and try and get an overall perspective on it. I've been thinking all the way through that it was a 2:1 and now I'm wondering if there's a possibility that it's a First. So I'm going to the Faculty of Arts assessment matrix…. (T7)

Norm referencing

I would have a look…and satisfy myself that the range of the marks…did seem to reflect what I'd written about the different pieces of work. So I'm saying 62 for the first one, 58 for the second one but conceivably they could be stretched with the upper one, 63 or 64 and the other one – possibly down. I don't think I would take it to 55, I would perhaps give it 56. (T5)

Concepts and Texts as Representations of Standards

- *assessment criteria, grade descriptors, statements of standards* and *marking schemes*: multiple terms used interchangeably , muddled as concepts
- Emphasis on ‘internalised’ standards:

  ‘internalised’, ‘absorbed’, ‘instinctively’, ‘got a sense of’, ‘in my mind’, ‘subliminal’, ‘rooted in my mind’, ‘got a mind set’, ‘implicit’, ‘have things in our heads’, ‘feel’, ‘familiar’ and ‘an understanding’.

‘Personal standards frameworks’

- Personalised lens for marking, internalised and loosely linked to explicit criteria, Learning outcomes, etc.

it’s a kind of you know almost subliminal level I’ve absorbed the outcomes and aims and I am using them. (T5)

……essentially the descriptions which exist in written documents which you've probably seen about what a First Class grade means, what a Second Class grade means and so on, they are rooted in my mind and have become part of my sort of experience really and I feel I can judge, I mean I could sit here and list all the criteria but there's no point in that. I feel I can judge now myself without referring to any kind of written standards but we do operate in accordance with those standards. (T5)

Individual differences in standards

- Trigger qualities
- Complexity of criteria: ‘It’s so multi-factorial you see’ (T1);
- Informal guidance points up differences to students:

the students have a success criteria grid and so according to that, and what I tell them, you know there are certain things they have to put in so there’s certain descriptive information that has to go in (T10).

Shared Standards

- Strong sense that standards are shared, if discipline specific:

There are things that are kind of implicit and in fact sometimes difficult to articulate but which nonetheless are relatively sound, that are disciplinary. They are just shared by being in the same discipline and provide a framework for marking that might not be available to other people outside that context. (T1)

Achieving the ‘correct’ mark?

I make the judgement on a piece of work but it’s always that niggling doubt. Am I right? And I can look at the criteria and think am I right? Without immediately going and giving it to someone else and asking what do you think? Which of course wouldn’t then be blind cross-marking anyway then it’s difficult to be sure. (T4)

I suppose with the rigorous second marking procedure and having the external examiner as well who looks at all our work so we have to be getting it right and it is quite a rigorous process really (T12)

Study 2: Aim of study

- To investigate the consistency of standards between assessors within and between disciplines.
- To investigate how their standards are shaped by their personal assessment histories, involvement in professional/disciplinary communities, experience of grading student work, and exposure to different universities and institutional and national reference points.

Methods

- 24 experienced assessors from 4 disciplines & 20 diverse UK universities;
- Each considered 5 borderline (2i/2.2 or B/C) examples of typical assignments for the discipline;
- Kelly’s Repertory Grid (1991 KRG) exercise used to elicit constructs that emerged from an in the moment evaluation based on actual student work – not idealised notions or marking guides.
- Followed by interview and *Social World Map* (Clarke 2005) exploring the influences on their standards

Experienced Assessor Research Project – KRG exercise construct sheet	
Name: EX19	University: New University
Discipline: History	Date: 2013

abc	X	cde	X	abe	X	bcd	X	ace	X	bde	X	acd	X	bce		ade		abd	

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		

Issues in using KRG to elicit standards

- KRG used elsewhere for its benefit in eliciting how expert examiners construe abstract demands;
- Relies on interpretation of constructs expressed;
- Numbers have weak meaning – signifiers of different comparative judgements;
- Assessors claimed to use the same standards as they would normally although examining usually involved different processes (e.g. seeing first markers’ mark);
- Assessors wanted more information about context (level, weighting, module info, had students been given feedback on drafts, etc.);
- Some claimed to find it difficult, others easy.

Range of constructs

- Between 3 – 10 (median 7) per assessor
- 37, 4 ‘surface’ constructs, 33 ‘global’

	psychology	nursing	chemistry	history
No. constructs	18	15	16	18
No. listed in criteria	7	5	No criteria provided	7

- Only 4 ‘extra’ constructs offered by 6 assessors

Consistency within disciplines

	psychology	nursing	chemistry	history
Constructs shared by 6 examiners			Quality of explanation	Historiograp-hy
Constructs shared by 5 examiners				Structure Academic Style,
Construct shared by 4 examiners	Use of evidence, Argument, Referencing, Academic style	Combined construct, analysis' wide reading English/ grammar, Referencing	Presentation /Legibility	Argument, Addresses the question, Wide reading

N.b. 13/37 constructs elicited by only 1 assessor.

Black = global construct Red = surface constructs

Disciplinary consistency within constructs

- assessors were asked to score each assignment on a count from 1 to 5 depending on how well it matched the construct identified (see example grid)
- This was designed to test the extent to which assessors judge work to be of a similar standard in relation to a specific quality, but this was hampered by lack of shared constructs.
- We investigated the 17 constructs used by at least 4 assessors in a subject discipline.

Disciplinary consistency within constructs

- In only 9 incidences out of a potential 85, all assessors within a subject area gave an essay roughly the same assessment for a particular construct (within two scores)
- only 2 examples where all the assessors award the same score for a construct.
- 42 instances (approximately half) where assessors rated the 5 different essays from 1 to 5 for the same construct

Consistency within constructs: psychology example

Assessors	1	2	3	4	5	6
Construct: Developing argument						
Essay A	5		2	5	4	
Essay B	2		2	1	1	
Essay C	1		5	5	5	
Essay D	3		3.5	2	6	
Essay E	4		1	5	1	

assessor 1: strong argumentation > weak argumentation
assessor 3: clear line of argument > really confused answer
assessor 4: develops an argument > no real understanding of argument
assessor 5: Tries to formulate an argument > Doesn't build an argument, answers in spurts (reversed)

Ranking of the different assignments

	Psychology essays				
assessors	A	B	C	D	E
1	4/5	1	2	3	4/5
2	5	2	4	1	3
3	4	2	5	3	1
4	3	2	4/5	1	4/5
5	4	2/3	1	5	2/3
6	no marks entered				
Range of rank	3rd- 5th	1st – 2/3	1st – 5th	1st – 5th	1st – 4/5

3 / 4 = joint 3rd/4th

Ranking of the different assignments

	Chemistry exam answers				
Assessors	A	B	C	D	E
1					
2	1	4/5	4/5	2/3	2/3
3	2/3	2/3	4/5	1	4/5
4	1	3/4	3/4	2	3/4
5	4	3	5	2	1
6	4/5	1/2	1/2	3	4/5
Range of rank	1st-5th	1/2- 4/5	1/2-5th	1st-3rd	1st-5th

3 / 4 = joint 3rd/4th

Social World Mapping: Findings

- Identified four locations where standards are seen to reside:
  - in explicit standards documents;
  - embedded in the individual - internalised;
  - in community processes*;
  - in student work.

* *Community Processes* refers to activities such as moderation, external examining or other disciplinary fora where the motivation is to discuss and calibrate standards.

Location of standards

- Most assessors located standards in documents or see them as internalised.
- Assessors conceive of community processes merely as a tool to check internalised standards or help in the interpretation of documented standards.
- Assessors rarely conceive of standards as located in student work.
- Some assessors were more reflexive about the provenance of their standards and their practices than others;
- Assessors commented 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.

Early conclusions

- a range of influences leads to different understandings of standards *(personal standards frameworks?);*
- listing criteria is only the first step in delivering consistent judgement;
- Assessors need awareness of variation in standards
- Assessors need greater engagement with explicit standards and to participate in greater discussion regarding the meaning of standards within the discipline (calibration Sadler 2013).

Developing quality assurance of higher education marking

- whilst a techno-rational perspective poorly represents the actual practice of standards in use, alternative, interpretivist accounts do not satisfy demands for reliability, transparency and fairness.

- Growing out of this research and much other work in the field including that from Cambridge assessment, there seem to me to be at least 4 dimensions where there is a mismatch between the messages of QA of assessment and the actual practice.

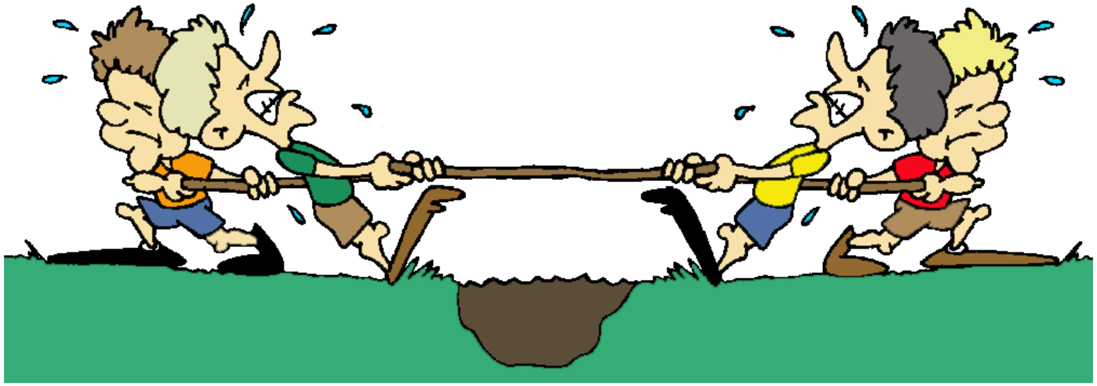
Standards in use - dimensions

Techno - rational	Hermeneutic/ socio-cultural
Explicit documentation of standards	Internalised, tacit standards
Criterion-referenced grading	Norm-referencing of judgement needed
Analytical judgement	Holistic professional judgement
Broad consensus on standards possible	Individualised standards or localised consensus

We lack QA (and advice for staff) which bridges this divide, which provides a firm basis for practice and which is understandable and credible to staff & students.

The tension

- **hermeneutic perspective**: the normalisation of bias, changing standards, inconsistency, norm referencing, and other features of professional judgement which generate concern.
- **the techno-rational approach**, with its beliefs in reliability and fixed standards, poorly represents the actual practice of grading in HE and isn't delivering consistency of standards.
- **Separation doesn't reflect reality?** Techno-rational QA processes do influence and give confidence to internalised, tacit judgement.

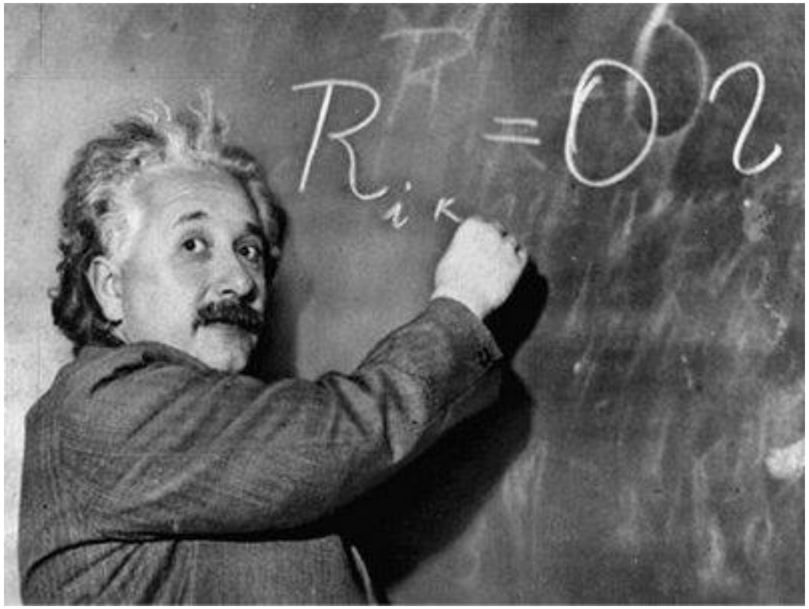


We need a quality model which:

- can be explained to all parties;
- has credence within the academy;
- which reflects or improves actual practice in a pragmatic way;
- bridges the limitations of explicit standards and the invisibility and variability of tacit standards to clearly demonstrate realistic and robust ways to achieve more effective security and fairness of standards in higher education.

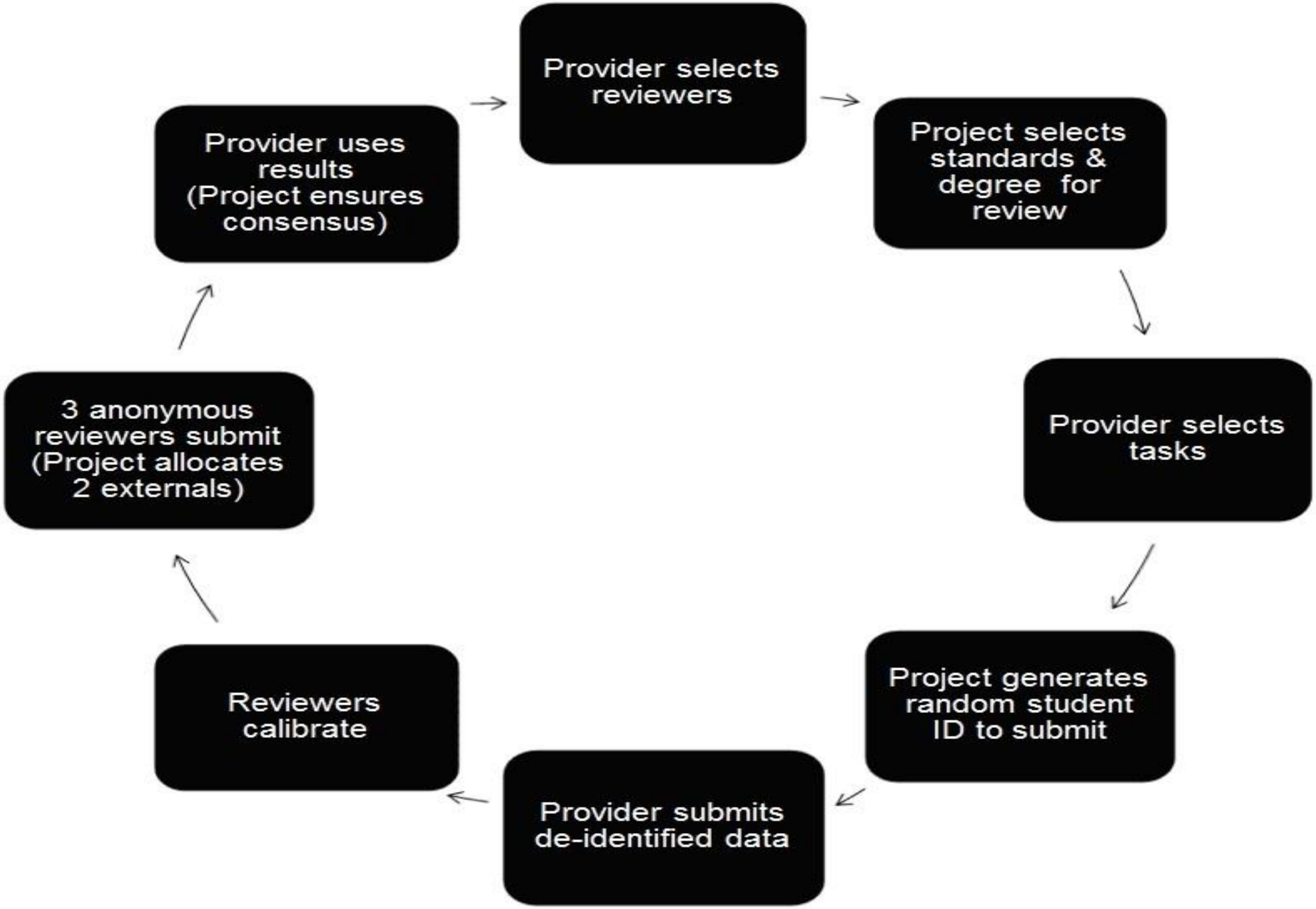
Possible elements

- Systematic and defendable processes for building consensus;
- Capitalising on the benefits of faculty creating and using explicit statements of standards;
- Ensuring new staff have a proper opportunity to engage with standards;
- Staff awareness and assessment literacy
- Student understanding of standards and grading



Better approaches to safeguarding standards in the assessment of complex work:

Achievement Matters



Provider selects reviewers

Project selects standards & degree for review

Provider selects tasks

Project generates random student ID to submit

Provider submits de-identified data

Reviewers calibrate

3 anonymous reviewers submit (Project allocates 2 externals)

Provider uses results (Project ensures consensus)

Conclusion

- Research is increasingly pointing up difficulties in assuring reliability and consensus of standards through higher education marking;
- Significant steps have been made to provide systems to assure quality and fairness;
- We need to interrogate those systems to consider how they can better match our growing understanding of professional judgement in general and marking practice in particular.

References and publications

Achievement matters http://achievementmatters.com.au/

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*, (iFirst).

Bloxham, S & Boyd P (2011) Accountability in grading student work: Securing academic standards in a 21st century quality assurance context. *British Educational Research Journal,* (2011) iFirst

Bloxham, S & Price, M (2013) External examining: fit for purpose? *Studies in Higher Education* (published on line 6th Sept)

Broad, B., 2003. *What We Really Value: Beyond rubrics in teaching and assessing writing*. Logan, Utah: Utah State University Press.

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

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

Delandshere, G. 2001. Implicit theories, unexamined assumptions and the status quo of educational assessment. *Assessment in Education* 8, no. 2: 113-133.

References con’t

Gipps, C. 1999. Socio-cultural aspects of assessment. *Review of Research in Education* 24: 355-392.

Kelly, G.A., 1991. *The psychology of personal constructs: Volume 1: A theory of personality*. London, UK: Routledge. (Original work published 1955)

Orr, S. & Bloxham, S. (2012) Making judgements about students making work: Lecturers’ assessment practices in art and design. *Arts and Humanities in Higher Education* (2012) (i-first)

Moss, P.A. and Schutz, A. 2001. Educational Standards, Assessment and the search for consensus. *American Educational Research Journal* 38, no. 1: 37-70.

Rust, C. et al (2005) A social constructivist assessment process model. *Assessment & Evaluation in Higher Education* 30 (3), 231-240

Sadler, D.R. 1987. Specifying and promulgating achievement standards. *Oxford Review of Education* 13, no. 2: 191-209.

Sadler, R.D., 2013. Assuring academic achievement standards: from moderation to calibration, *Assessment in Education: Principles, Policy and Practice,* 20(1), pp. 5-19.