4 November 2015: 07:00AM GMT

The multiple limitations of assessment criteria

Em. Prof. Sue Bloxham (University of Cumbria, UK)

Special preview session **aheconference.com**

Your Webinar Hosts

Professor Geoff Crisp,
Dean Learning and Teaching, RMIT University
geoffrey.crisp[at]rmit.edu.au

Dr Mathew Hillier,
Office of the Vice-Provost Learning & Teaching,
Monash University
mathew.hillier[at]monash.edu

Just to let you know:
By participating in the webinar you acknowledge and agree that:
The session may be recorded, including voice and text chat communications (a recording indicator is shown inside the webinar room when this is the case).
We may release recordings freely to the public which become part of the public record.
We may use session recordings for quality improvement, or as part of further research and publications.

e-Assessment SIG
ascilite
SHAPING THE FUTURE OF TERTIARY EDUCATION

Transforming Assessment webinar series:

The multiple limitations of assessment criteria

Sue Bloxham

UNIVERSITY of Cumbria

Webinar discussion

- I will build in some discussion points during the webinar, but
- Please feel free to ask questions or make comments at any point.

Assessment criteria: Transparent and fair marking?

- Criteria are designed to make the processes and judgements of assessment more transparent to staff and students and to reduce the arbitrariness of staff decisions (Sadler 2009).
- ‘production, publication and discussion of clear assessment criteria ….[is now regarded as] a sine qua non of an effective assessment strategy’ (Woolf 2004: 479)

Aim of seminar

- The aim of this webinar is to draw on research to explore the use of assessment criteria by experienced markers and discuss the implications for fairness, standards and guidance to students.

Evidence of inconsistency

- poor reliability and consistency of standards amongst those assessing complex performance at higher education level
- Many studies:

  O'Hagan & Wigglesworth, 2014; Hugh-Jones et al 2009; Read *et al.*, 2005; Price, 2005; Shay 2004; Baume *et al.*, 2004, Norton *et al.*, 2004, Elander and Hardman 2002; Leach *et al.*, 2001; Wolf 1995; Laming 1990

Causes of inconsistency

- Different professional knowledge, experience, values,
  (Read et al, 2005, Smith & Coombe, 2006);
- Marking habits (Wolf 1995) & 'standards frameworks' (Bloxham et al 2011)
- Different expectations of standards at different grade levels
  (Grainger, Purnell, and Zipf 2008; Hand and Clewes 2000).
- Ignoring or choosing not to use the criteria
  (Price & Rust, 1999; Ecclestone, 2001; Baume *et al.*, 2004; Smith & Coombe, 2006);
- Different interpretation of criteria or standards
  (Webster, et al, 2000; Moss & Schutz, 2001).
- Use personal criteria different to those stated.
  (Broad 2003; Dobson, 2008; Greatorex, 2000; Hawe, 2002; Baume *et al.*, 2004; Price, 2005; Read *et al.*, 2005, Webster, Pepper & Jenkins, 2000)
- Importance given to different criteria
  (Read *et al.*, 2005; Smith & Coombe, 2006);
- Focus on different aspects of student work
  (O'Hagan & Wigglesworth, 2014).

Discussion

- What is your experience of reliability and standards in marking?
- Have assessment criteria been helpful?

Study

Part of wider project on standards in use by experienced (external) examiners.

One aim (reported in this paper):

- To investigate the consistency of standards between examiners within and between disciplines.

Collaborators in research

- Margaret Price
- Jane Hudson
- Birgit den Outer

All from the ASKe Pedagogical research centre, Oxford Brookes University, UK

OXFORD BROOKES UNIVERSITY

Method

- 24 experienced examiners 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.

External Examiner Research Project – KRG exercise construct sheet

Name: EX1									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		

Consistency of judgement: ranking the assignments

	Assignments				
	A	B	C	D	E
psychology	3rd-5th	1st – joint 2nd/3rd	1st-5th	1st-5th	1st – joint 4th/5th
nursing	1st- joint 3rd/4th	1st-5th	Joint 1st/ 2nd – 5th	Joint 1st/ 2nd – 4th	1st – joint 3rd/4th
chemistry	1st-5th	Joint 1st/ 2nd – Joint 4th/5th	Joint 1st / 2nd – 5th	1st – 3rd	1st-5th
history	Joint 1st / 2nd – 3rd	Joint 1st / 2nd – 4th	Joint 2nd/ 3rd – 5th	Joint 2nd/ 3rd – 5th	1st – Joint 1st/2nd

Case analysis: history

	A	B	C		E
Range of marks for each essay	1st – 2.2 (A-C)	1st – 2.2 (A-C)	2.1. – 3rd (B-D)	2.1. – 3rd (B-D)	1st – 2.1 (A-B)
Range of rank for each essay	J1/2-3rd	J1/2-4th	J2/3-5th	J2/3 -5th	J1-J1/2

Range of marks for each assessor	
1	1st-2.2/2.1 (A – B/C)
2	1st-3rd (A – D)
3	Mid 2.1-Low2.2. (B – C)
4	1st-3rd (A – D)
5	2.1.-2.2 (B – C)
6	1st-3rd (A – D)

Red = less use of full range of marks

The role of constructs

- overall agreement on a mark by assessors appears to mask considerable variability in individual criteria;
- The difference in the historians' appraisal of individual constructs was further investigated and five potential reasons were identified that link judgement about specific elements of assignments to potential variation in grading.

Reason 1:
Using different criteria to those published

- Difficult title/ question attempted
- Good attempts to define constructs
- Attempts to set up essay with introductory paragraph
- Understanding of wider context
- Quality of explanation (includes diagrams to explain/underpin answers) and sufficient detail
- English/ grammar/ proof reading
- Referencing/ citation

- Analysis/ critical analysis
- Addresses the question
- Structure/ organisation
- Good conclusion
- Style/ Academic style/ register
- Presentation/ legibility
- Historiography
- Wide reading,
- Depth/ quality of Knowledge
- Developing argument, argumentation
- Use of theory

Reason 2: Assessors have different understanding of shared criteria

	Construct: engagement with historiography				
Assessors	Essay A	Essay B	Essay C	Essay D	Essay E
1	2	1	3	5	1
2	4	4	2	5	1
3	3	4	1	1	5
4	4	2	5	5	1
5	3	3	1	3	5
6	5	5	6	5	1

1. *Engages well with historiography > Hardly engages with historiography (reversed)*
2. *Historiographically determined > Less determined by historiography*
3. *Engagement with historiography > Unawareness of historiography*
4. *Awareness of historical debate, historiography > Absence of debate*
5. *Clear investigation of previous arguments in the area > Not enough use of historiography*
6. *Engages with historiography > Doesn't explicitly discuss the historiography*

Consistency within constructs:

	Construct: Developing argument, argumentation				
Assessors	Essay A	Essay B	Essay C	Essay D	Essay E
1	1	2	5	4	3
2	didn't use construct				
3	1	2	5	4	-1
4	1	2	5	4	-1
5	5	3	1	2	3
6	didn't use construct				

1. *Argument excellent > argument adequate*
3. *Argument focus > narrative focus (reversed)*
4. *Reasonable argument > superficial argument*
5. *Clear exposition of argument > contradiction of argument.*

Reason 3. Assessors have a different sense of appropriate standards for each criterion

	Construct: Developing argument, argumentation				
Assessors	A	B	C	D	E
1 *Argument excellent*	1	2	5	4	3
3 *Argument focus*	1	2	5	4	-1
4 *Reasonable argument*	1	2	5	4	-1

Reason 4. The constructs/criteria are complex in themselves, even comprising various sub-criteria which are hidden to view

	Construct: structure				
Assessors	Essay A	Essay B	Essay C	Essay D	Essay E
1	1	2	2	3	5
2	1 2	2 5	5 5	1.5 2	0 1
3	2	3	8	5	1
4	4	2	5	5	1
6	1	4	5	5	1

4. *Extremely well-structured > not so well structured*
6. *Clear structure and signposting > jumps in with no signposting*

N.B. Assessor 5 did not use this construct

Construct language: structure

1	*Keeps a logical and analytical structure all the way through*	>	*Loose structure*
2.a	*Thematically and analytically structured*	>	*Narrative dominated by chronological approach*
2.b	*Balanced in level of attention to all structural components*	>	*Imbalanced in level of attention to all structural components*
3	*Effective structure*	>	*Weak structure*
4	*Extremely well-structured*	>	*Not so well structured*
6	*Clear structure and signposting*	>	*Jumps in with no signposting*

Variation may be a feature of our methodology but similar confusion is likely to exist in criteria simply described

Reason 5. Assessors value and weight criteria differently in their judgements

- Only consistency in ranking related to lower ranks for surface constructs
- In the constructs which they largely shared such as *structure*, the rankings ranged between 1 and 5, and between 2 and 5 for *historiography.* The ranking for *style/ academic style* ranged between 1 and 10.

Discussion

- Questions about the research?
- Should we recognise the impossibility of a ‘right’ mark in the case of complex assignments?
- What are the implications for fairness, standards and guidance to students?

Social processes: ‘Flipping’ the assessment cycle?

- Emphasise pre-teaching moderation – Discussion of assessment tasks, criteria and exemplars could:
  - Improve assessment design
  - Inform teaching (assessment for learning)
  - Inform dialogue with students (reduce teacher inconsistency)
  - Help develop shared standards amongst teaching team
  - Improve consistency of marking judgements
  - Allow for more discussion (calibration) of standards because it can take place without the time pressure on moderation at the end of a course
- End of course moderation can then focus on what is important (very high stakes and borderline work) – do a little moderation well rather than a lot superficially

Social processes – emerging in quality requirements

For example:

‘practices which promote and support consistency of marking by and between staff, including dialogues which enable a shared understanding of standards’ (QAA Quality code, chap 6, p13)

Example: Calibration of standards

Achievement matters: accountancy in Australian universities

- Academics from all types of higher education institutions took part in 'calibration' activities, independently rating both the validity of assessment tasks and examples of final year student work and then meeting to discuss and agree the judgements.
- These academics then participated in the anonymous review of assessments (briefs and student work) from other providers.
- The external calibration of discipline standards resulted in a measurable decrease in variability in academics judgements.

Learning standards the way tutors learn?

- emphasises holistic judgement processes;
- is embodied in real judgements;
- is dialogical;

- it takes place over time; recognising that standards cannot be acquired in one attempt;
- Recognises the nature of complex judgement and the context for University assessment.
- Encourages a view of knowledge as contestable

(Bell et al 2013)

Conclusions

- Study provides empirical support for previous research and theoretical ideas; 'who marks your essay becomes more important than the content of the essay itself.' (O'Hagan & Wigglesworth, 2014)
- Five reasons likely to work in combination;
- More detailed criteria is not the answer;
- Social processes?
- Don't strive for 'right' marks? – perhaps greater fairness and accuracy emerges from multiple assessors and assessment opportunities providing several judgements on individual students';
- Share nature of professional judgement with students?

References

Baume, D., et.al. (2004) What is happening when we assess, and how can we use our understanding of this to improve assessment? *Assessment and Evaluation in Higher Education,* 29 (4), 451-477.

Bell, A., R. Mladenovic, and M. Price. 2013. "Students' perceptions of the usefulness of marking guides, grade descriptors and annotated exemplars." *Assessment & Evaluation in Higher Education* 38 (7): 769-788. Doi:10.1080/02602938.2012.714738

Bloxham, S., Boyd, P. and Orr, S. 2011. Mark my words: the role of assessment criteria in UK higher education grading practices. *Studies in Higher Education* 36, no. 6: 655-670.

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

Dobson, S., 2008, 'Applying a validity argument model to three examples of the viva', *Nordisk Pedagogik,* 28, pp. 332–44.

Elander, J. and Hardman, D. 2002. An application of judgement analysis to examination marking in psychology. *British Journal of Psychology* 93: 303-328.

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

Greatorex, J., 2000, 'Is the glass half full or half empty? What examiners really think of candidates' achievement', paper presented at the *British Educational Research Association Conference,* 7–10 September 2000, Cardiff.

Hand, L. and Clewes, D. (2000) Marking the Difference: An Investigation of the Criteria Used for Assessing Undergraduate Dissertations in a Business School, *Assessment and Evaluation in Higher Education,* 25 (1):5-21.

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

Hugh-Jones S., Waterman MG., Wilson I. (2009) Accessing and understanding the tacit dimensions of assessment. *Psychology Learning & Teaching*, **8** (2), pp. 7-15.

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

Laming, D. 1990. "The Reliability of a Certain University Examination Compared with the Precision of Absolute Judgements." Quarterly Journal of Experimental Psychology 42A (2): 239—54.

Leach, L., Neutze, G. & Zepke, N., 2001, 'Assessment and empowerment: some critical questions', *Assessment and Evaluation in Higher Education,* 26(4), pp. 293–305.

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

Norton, L., et.al. (2004) *Supporting diversity and inclusivity through writing workshops. Paper presented to the International Improving Student Learning Symposium, Birmingham, UK, 6–8th September.*

O'Hagan, S.R & Wigglesworth, G (2014) Who's marking my essay? The assessment of non-native-speaker and native-speaker undergraduate essays in an Australian higher education context, *Studies in Higher Education*, DOI: 10.1080/03075079.2014.896890 [accessed 15th June 2014]

Price, M. (2005) Assessment Standards: The Role of Communities of Practice and the Scholarship of Assessment, *Assessment and Evaluation in Higher Education,* 30 (3), 215-230.

Price, M. and Rust, C. (1999) The Experience of Introducing a Common Criteria Assessment Grid Across an Academic Department, *Quality in Higher Education,* 5 (2):133-144.

Read, B., Francis, B and Robson, J. (2005) Gender, bias, assessment and feedback: analyzing the written assessment of undergraduate history essays, *Assessment and Evaluation in Higher Education,* 30 (3):241-260.

Sadler, D.R. 2009, Indeterminancy in the use of preset criteria for assessment and grading, *Assessment & Evaluation in Higher Education,* 34, no. 2: 159-179.

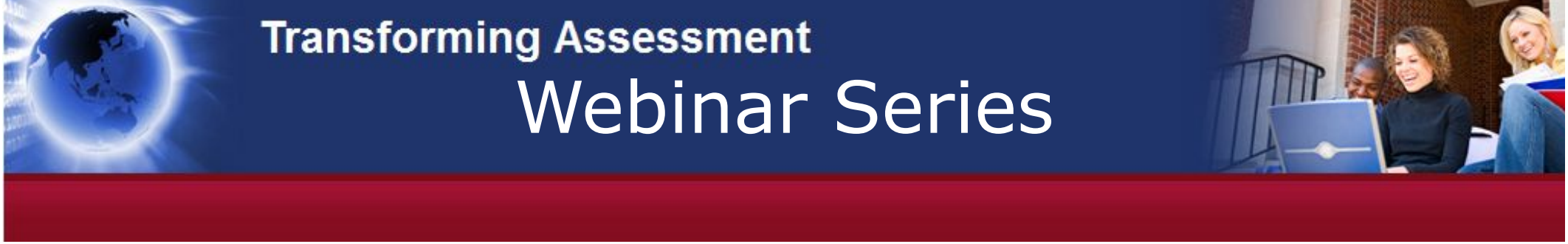
Smith, E. and Coombe, K. (2006) Quality and qualms in the marking of university assignments by sessional staff: an exploratory study, *Higher Education,* 51 (1):45-69.

Webster, F., D. Pepper, and A. Jenkins. 2000. "Assessing the Undergraduate Dissertation."

Assessment and Evaluation in Higher Education 25 (1): 71–80.

Wolf, A. 1995. *Competence-based assessment.* Buckingham: Open University Press.

Woolf, H. 2004, Assessment criteria: reflections on current practices, *Assessment and Evaluation in Higher Education,* 29, no. 4: 479-493.



Session Feedback Survey

With thanks from your hosts

Professor Geoff Crisp,
Dean Learning and Teaching, RMIT University
geoffrey.crisp[at]rmit.edu.au

Dr Mathew Hillier,
Institute Teaching and Learning Innovation,
University of Queensland
mathew.hillier[at]uq.edu.au

Recording available
http://transformingassessment.com

e-Assessment SIG
ascilite
SHAPING THE FUTURE OF TERTIARY EDUCATION