

13 Assessing creativity through consensual assessment

Tom Balchin

Introduction

The aspect of creativity that poses the greatest challenge to higher-education teachers is how to assess and evaluate it. This chapter considers some of the problems associated with the assessment of creativity in higher-education students' learning, and describes a method for assessing creative performance that has been proven to work in school education and which could be adapted to higher education.

In synthesising the results of Imaginative Curriculum discussions on assessing creativity, Jackson (2005d) commented that, while many teachers believe that it is possible to help students use their creative abilities to better effect, far fewer think it is possible to assess these capabilities reliably – and even fewer are prepared to try to do it. Higher-education teachers hold diverse views on whether creativity can be assessed. A minority of teachers believe that students' creativity is evaluated in specific components of a programme through explicit assessment criteria. A majority of teachers believe that insufficient attention is given to recognising students' creativity and that, at best, evaluation and recognition is haphazard and implicit – a bi-product of assessing higher-order thinking skills. A third group of teachers value creativity but don't know how to assess it. A fourth group believe that it is just not possible to assess creativity or, if it was, it would be so subjective as to be meaningless or require too much effort. The final group believe that the very act of assessing creativity will cause it to disappear! But we should not be put off by such a contradictory and disparaging list of beliefs, as there is a sense that most higher-education teachers would assess creativity in students' higher-education learning if they were given appropriate support (mainly time to change), guidance and cultural encouragement to do so.

What is also clear is that many teachers believe that some forms of summative assessment are major inhibitors of creativity. Higher education in the UK is now based on an outcomes model of learning in which teachers attempt to predict the outcomes from a process that they orchestrate. However, learning emerges from creative processes in unpredictable ways – and, unless the learning outcomes, assessment criteria and assessment methods accommodate this way of thinking, it is unlikely that a student's creativity can be encouraged,

demonstrated and evaluated through the assessment process. A further problem with the common forms of summative assessment in higher education is that they generally do not permit failure (a distinct likelihood in high-risk situations where students are attempting to do radical things for the first time). Summative assessment generally encourages students to play safe, and to seek to achieve the outcomes intended by the teacher, rather than the outcomes the student may like to achieve given more flexibility.

Manifestations of creativity in students' learning

Creativity is manifested in the engagements of learners in creative processes and in the products that result from their creative endeavours. In higher education, products might include such things as: essays and many other forms of writing, including reports, diaries and reflective logs, poems, the products of electronic discussions; posters, the results of problem working, design and synthesis – like independent projects, laboratory or field notebooks (Jackson, 2005d). They include visual and graphical representations – designs, sketches, drawings, paintings, photographs, videos, computer animations; physical and virtual models and constructions; performance – theatre, role play, simulation, dance, song and live or recorded presentations, and many other things.

Similarly, there are many possibilities for processes and contexts within which creative acts might occur and products be created. Processes may be individually constructed and self-directed or be constructed collaboratively with other people or involve elements of both. They may be developed within a pedagogy – like enquiry or problem-based learning, role play, design process or independent or team project, and a particular context like fieldwork or work placement. They may constitute rehearsals directed to creative performance, as in music or dance. Processes and creative journeys can be evaluated through direct observation of students, through video or audio records or through diaries and reflective personal accounts supported by evidence that authenticates the account.

Assessing creativity is difficult

Much effort has been devoted (from the end of 1950 onwards) to the identification of the special characteristics that lead to creativity, and to the ways that creative people perform their creative acts. This effort has led to the development of numerous instruments that are used in the identification of creative talent. Kaltsounis and Honeywell (1980), for example, have published an exhaustive list of 'creativity tests' and Torrance and Goff (1990) have identified 255 such instruments.

It must be emphasised that different researchers have studied different aspects of creativity and that often these cross-purposes have been reflected in the measuring processes. There is not one *accepted* method for the measurement of creativity in individuals or groups working collectively and by using different techniques (tests of cognition, attitudes, interests, personality, biography, etc.) to

Evaluating creativity through assessment 175

1 assess creativity, researchers are actually studying different phenomena!
2 Another facet of such tests (like the Torrance tests of creative thinking) is that,
3 although they have been shown to have a high validity rating, their relevance to
4 university settings, where teachers need to recognise and reward students' cre-
5 ativity in the specific contexts in which they are learning, is questionable.

6 Assessing students' creativity is not a simple matter, so there has to be a good
7 set of reasons to persuade teachers to do so. Here are some reasons for assessing
8 creativity:

- 9
- 10 • It helps to demystify and operationalise the idea.
- 11 • It helps to develop a language for communication among professionals
- 12 about the nature of creative abilities, skills and potential in their disciplinary
- 13 contexts.
- 14 • It helps teachers to engage students and helps students to evaluate and
- 15 recognise their own strengths and talents, enabling them to understand
- 16 themselves better. By focusing students' attention on learning through
- 17 assessment, they are more likely to believe that their creative enterprise is
- 18 valued and the intrinsic motivations that are associated with creativity are
- 19 more likely to be harnessed.
- 20 • By involving students in the assessment process, they are developing the
- 21 skills of self-evaluation and self-judgement that are crucial to further learn-
- 22 ing and the improvement of their own performance and creative potential
- 23 (Cowan, this volume, Chapter 12).
- 24 • It can encourage students to take risks in order to learn. Tolerating the non-
- 25 accomplishment of desired goals will encourage students to engage in
- 26 riskier endeavours.
- 27 • It provides feedback to teachers to enable them to discover unrecognised or
- 28 untapped potential/talent.
- 29 • It provides information to help teachers evaluate the effectiveness of their
- 30 strategies for promoting students' learning, and enables them to construct a
- 31 picture of students' understandings of their own creativity.
- 32

33 In the following section I describe the idea of consensual assessment and
34 propose that the approach offers a useful and relevant route to evaluating stu-
35 dents' creative efforts. Furthermore, the process of developing consensual
36 assessment within peer groups of teachers and students can be a powerful source
37 of personal or professional development.

38 **Consensual assessment**

39
40 This concerns the idea that a product or response is creative to the extent that
41 appropriate observers consensually agree it is creative. Appropriate observers
42 are those familiar with the domain in which the product was created or the
43 response articulated. This is not a particularly profound proposition, but the
44 technique can be shown to be more reliable than many other tests, probably
45 because it involves the views of experts from the domain the creative work

under examination resides in. Teachers, who can be considered as experts in a subject domain, can recognise how a contribution from a student may vary from established works, constructs or the thinking or performances of other students. The most important criterion for creativity assessment procedures is that any ratings produced should be found to be as *reliable* as possible (Balchin, 2005). The discursive process used is not a weakness in the approach; inter-judge reliabilities reported in this area of research have been found to be moderately high.

To discover the reliability of this technique in relation to others, I examined more than 100 examples of creativity measurements from many diverse fields, from Jackson and Messik's (1965) tests of creative potential, to the Amabile (1995) KEYS (assessing the climate for creativity) Inventory and found that a few of these approaches require small groups of people (i.e. supervisors, peers, teachers) to make judgements about products, ideas or other people after discussion.

Factors to be considered

The first factor to consider is who the judges should be and what the judges should be looking for. In schools, it is the teachers, perhaps with the help of field-based professionals, who are the judges. The experiment in using consensual assessment described below was predicated on the basis of teachers being the judges of students' creativity. But in higher education, the situation is more complex. Teachers or small teaching teams are responsible for students' learning and the outcomes of their learning experiences, but the students themselves are also responsible for engaging with learning processes in ways that optimise their learning and achievements. There is a level of learner autonomy and expectation of independence in higher education that is not expected in schools. In the absence of a universal discourse on creativity, the issue for higher education is how to help students to recognise and make claims for their own creativity (Cowan, this volume, Chapter 12). Helping students to evaluate their own creativity must go hand-in-hand with helping them to make claims. Teachers and students should form an appropriate community of learners within which understandings and meanings of creativity can be socially constructed and evaluated in the precise situations and contexts for learning. Consensual assessment in higher education might therefore involve both the teacher and the learners.

An important question for those evaluating creativity is whether evaluators are able to distinguish creativity from other constructs such as intelligence, achievement, competence, etc. A non-systematic review of some appropriate studies indicates otherwise. Holland (1959) reports an interesting study in which school teachers, heads and guidance counsellors rate students on 12 traits, including originality. The results are revealing; originality correlated 0.72 with speaking skills and 0.84 with writing skills. The other nine correlations that involved originality ranged from 0.50 to 0.65. This indicates that there is a general difficulty in discriminating creativity from other attributes. Alternatively, perhaps these attributes are integral to an individual's creativity and are bound up in any judgement.

1 People who participate in the process of consensual assessment also have to
2 be wary of establishing opinion-sets from (or before) the outset, which influence
3 *all* of their judgements. Hocevar and Bachelor (1989) suggest that this may be
4 termed the 'halo effect'. It is a dangerous feature of small groups that could
5 result in a failure to discriminate creativity. Interestingly, Surowiecki (2005)
6 found the same phenomenon in research that indicates groups (supplied with
7 knowledge about a situation or product with a similar few pieces of knowledge
8 in common) tend to discuss only the information in common that they view as
9 most important. Rossman and Gollob knew about this effect in 1975, and stated
10 that groups make better judgements regarding creativity and intelligence when
11 the judgements are based on diverse information.

12 Törrönen (2001) indicates that small groups can polarise views, leading to the
13 more extreme views of group being put forward as solutions. Group polarisation
14 is still a phenomenon that is not well understood. It has been found to be the
15 result of people doing their best to figure out what the right answer is and, in
16 doing so, inadvertently makes people more likely to advocate extreme positions:
17 which sociologists have termed the 'risky shift'. This seems to argue for larger
18 groups and communities of interest rather than small, introverted groups. But
19 solid as the evidence for polarisation is, so too is the evidence demonstrating
20 that groups can become, as it were, depolarised, and that small groups make
21 better decisions and come up with better answers than most of their members,
22 and, surprisingly often, the group outperforms even its best member (Sunstein,
23 2003).

24 So, despite the amount of literature on the measurement of creativity, it
25 seems that engaging in the process of understanding and evaluating creative
26 achievement and activities through processes of consensus building to reach
27 agreement is likely to yield real benefits to the professional understandings of
28 teachers and, if students are involved, to students' understanding of their own
29 creative abilities.

31 **An example of consensual assessment**

32 A research study to design creativity assessment tools for the field of design and
33 technology education (hereafter d&t) was undertaken in 2004 with 14 London
34 secondary school Heads of Department. The aim was to gather d&t expert
35 teachers in their very practical field to see if they could use a creativity feedback
36 tool (product sheet) using consensual assessment. The detailed results of the
37 research can be found in Balchin (2005); what is important here are the lessons
38 and general principles that were derived from the exercise.

39 The building of a *creativity feedback package* (CFP) was initiated by finding
40 out the extent to which existing models of creative behaviour interact in the
41 appropriate diagnoses of creative performance. It was found that it was possible
42 to transfer a contextual organiser for creativity known as the '4 Ps' – people,
43 process, product, and press (climate) – directly to the field of d&t. Rhodes
44 (1961) first coined this phrase, but I have found Murdock and Puccio's (1993)
45 thesis concerning its meaning to be more relevant to creativity assessment. Their

finding was that *it is not preferable or even possible to incorporate all four facets during assessments*. In other words, it is not essential to know everything about the creation and situation; but, by gathering data about a combination of necessary elements (normally three out of the four Ps), a reasonably efficient evaluation or prediction of creative activity can be produced.

The *creative product* feedback sheet (one of three feedback sheets within the CFP) was designed as a score card to guide the teacher in judging and marking a product for creativity. The sheet has nine criteria for products to be scored against. However, numbers were not critical to this exercise, because the judge will have his/her unique scale of judgement and frames of reference to the product. The scores therefore have no real meaning as numbers, they only help the scorer get to grips with the criteria and force judgements (Balchin, 2005).

Four criteria describe the creative *concept*, or idea, and three criteria describe the *quality of build*, which evaluate how well the creative thoughts have been shown in the product. The emphasis in this evaluation sheet was that creativity is seen in *both* the concept *and* the standard of build that the product showed. But it is the *concept* stage where the unique ideas are brought forth, and the product sheet is the manifestation of those creative ideas. The latter cannot occur without the former; I propose that the *quality of build* is a 'vehicle' for the creative thought. The scorer therefore is given a 12-point Likert scale to use, intended to mirror the A, B, C and D of traditional mark schemes, with the corresponding pluses and minuses.

These criteria are: *uniqueness, associations of ideas, risk-taking, potential, operability, well-craftedness and attractiveness*. Concise definitions were attached to these criteria, which were themselves subsumed from an exhaustive list of qualities of creative products. The criteria were trialled in schools for nine months in order to see if they were able to describe creativity simply and without the danger of overlapping of criteria, yet retaining the complexity of the phenomenon.

The pilot study involved 14 heads of department (one panel of five, and three panels of three) as expert judges. The judges were each asked to use the criteria on the CFP product page to produce an *individual*, then a *consensual*, score of ten different products. The judges were all given the individual product sheet, asked to choose any of the ten products they liked, and given 30 minutes to assess them according to the criteria on the product sheet. The one rule was that if more than one judge was working on the same product, no collaboration would be allowed. It is important to bear in mind that no accompanying portfolio/coursework was present. The judgements were made from the products alone, to see if agreement over the criteria could occur even when teachers did not know anything about the task or making process.

After this individual scoring time, the judges were asked to split into three groups and review three out of the ten products. The judges were split into three groups and given 40 minutes to rotate around the products, debating and reaching consensus before scoring each criteria (1–12). Serious debate was seen to take place inter-group. However it was observed that the judges were able to agree about criteria, without having to compromise their initial thoughts too

1 much. In this way, individual scores were produced for ten products and consen-
2 sual scores from three groups were produced for three products.

3 The results are reproduced below (using those produced by evaluating one of
4 the products as an example) for the seven criteria used, with the degree of spread
5 between the scores given in each criteria marked with circles. The horizontal
6 line that many are on indicates the lateral degree of spread of each of the scores
7 between '1' (furthest left) and '12' (furthest right). The spread of scores are indi-
8 cated in the column of numbers to the right. The significance of the thin vertical
9 line will become clear when this spread of scores are compared to those that
10 resulted when products were scored consensually.

11 The chart shows the result (spread of scores) from the individual consensual
12 assessing. Only one of the seven criteria, *association of ideas*, fell outside a
13 spread of four marks. This was a difficult definition to present to teachers, but
14 interestingly, two out of three scored it in exactly the same way. The other was
15 working from her own particular *frame of reference*, which is a clue to the value
16 of consensual assessing, which has the power to take all frames of reference into
17 account.

18 Taken as a whole, the judges (scoring individually) showed a high degree of
19 consensus over each score. This indicated that (most of) the criteria and defini-
20 tions worked, and that there was enough 'clear space' between each definition
21 for each to make sense. It also showed that the judges were pretty good at recog-
22 nising the inherent creativity of a product without any discussion, simply by ref-
23 erence to other similar products that they are aware of. Moreover, individuals
24 can arrive at similar conclusions without viewing any of the designing and
25 making of the product.

26 Did consensual scoring tend to diminish personal bias or dislike, to give a
27 fairer impression of creativity occurring? This was difficult to prove. However,
28 when scoring consensually, it can be seen that the mean spread of scores per cri-
29 teria diminished significantly when the judges were allowed to sharpen and hone
30 their professional judgements about the product through debate aimed at reach-
31 ing agreement.

32 The thin vertical lines on both Figures 13.1 and 13.2 show a *mid-line* between
33 the scores. It brings to the attention the way that the consensual scoring groups
34 tended to score the product more positively in terms of creativity; a feature seen
35 on each consensual assessment when compared to individuals assessing the
36 same product.

37 38 **Comparing scores for individual and consensual assessment**

39 The first four criteria are views about the product's *concept*. All four were
40 scored very close to each other. For example, when it came to scoring *associ-*
41 *ation of ideas*, each consensual group of four to six judges scored it *exactly the*
42 *same* (without conferring between groups, as per instructions for the whole exer-
43 cise). This was a significant and positive finding, because the product was unfin-
44 ished. Nevertheless, the judges were able to see through the unfinished state and
45 glimpse the creative thoughts that caused the construction to be the way it was.

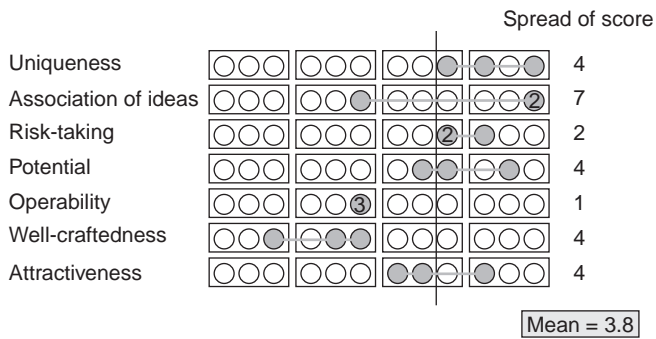


Figure 13.1 Spread of scores from individual scorers of one of the products.

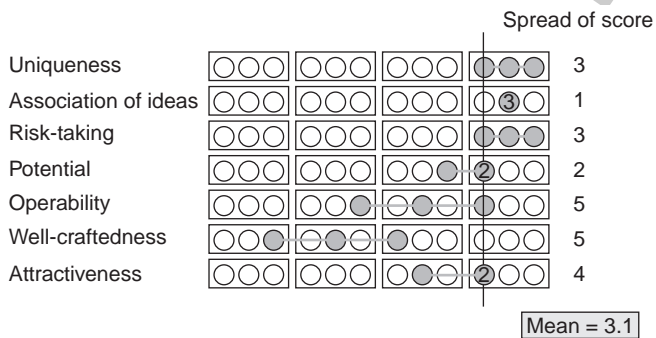


Figure 13.2 Spread of scores from consensual scorers of the product.

The unanimous agreement about this particular criteria (produced by three groups of judges, each scoring consensually) held fast with the other two products under group assessment. The *association of ideas* is a most important criteria at the heart of creativity, and it was important that each group of consensual assessors were able (with a one-line definition) to understand the notion well.

The last three criteria concerned the product's *quality of make*. In general, if the condition of the item is poor, creativity is harder to see and score. This is why *operability*, *well-craftedness* and *attractiveness* are important markers in creativity judgements, because they relate to the feeling of rightness of a product and the sense that ideas have been pulled together into a coherent whole. It can be seen that the three individual judges, and the three consensual judging groups especially, dispersed their scores for the product's *quality of make* around the scoring box, from 3–10. This was in contrast to the dispersal of scores for the product's *concept*, where the scores ranged from 8–12 (except for one score given to *association of ideas* by one group). When the judges were asked about this disparity, the answer was that the product was difficult to judge in terms of

operability because there were different opinions over the intentions of the students who made it. This was an unforeseen factor, but it did highlight some of the problems of group assessing without knowing what the students' tasks were all about.

The judges were also given aspects of creativity to think about in relation to the product in front of them. Free-response boxes were provided with *prompts* to help them to reach their individual scores.

- What kind of unity can you see in it?
- What sense of rightness do you get from this product?
- To what extent does it give an understated or economical solution?
- Is it a coherent whole?
- What kind of aesthetic sense of organisation can you see in it?
- How far does it go to resolving a discord?

The result – a compilation of judges' views – was found to be useful because it demonstrated how different judges can make parallel judgements as they scored products in terms of creativity. The judges seemed to be able to converge on scores to give to the concepts, even though their ideas about the standard of build that should be achieved for creativity adequately to be shown differed. This was an indicator that creativity can be *reliably* recognised in a product, even if it was badly made and 'didn't work'.

This was a small sample of judges and products, but it illustrates a process for developing and refining judgements about creativity that seems to work. The judges agreed that the creativity resides in the concepts or idea. They could see that a product which may score badly in terms of *well-craftedness*, *operability* and *attractiveness*, many still score very highly in terms of *uniqueness*, *association of ideas*, *risk-taking* and especially, *potential*. They believed that it could encourage more creative thought during designing stages, and begin to give some sort of feedback/recognition to creative 'failure'.

Adapting this approach to higher education

One of the benefits of consensual assessment is that it engages teachers in purposeful professional discussion: in this case, discussion about creativity in relation to the products that their students produce within the learning processes and activities that they create. Through such conversations, teachers develop new and deeper understandings about the nature of creativity in the particular contexts in which it is being developed and displayed. Ultimately it is a means of developing better understanding of creativity within the disciplinary field.

The criteria to work around are the key to this process, however, as they provide a starting point for the development of mutual understandings about creativity. These might be developed through the discursive process involving only teachers, or they might be part of a process of discussion and negotiation involving teachers and students.

182 Tom Balchin

The kind of consensual assessment sheet outlined above can be adapted not only to situations that require products to be produced by students, but also to situations where outcomes other than products are sought. For instance, in the solution of problems in quantum physics or a performance in dance, appropriate criteria may substitute for the *quality of build* criteria. The prompts for discussion are easily adaptable to productions in all disciplinary fields. The sheet itself is simple enough to be completed by a group of judges in approximately ten minutes.

In the author's view, *summative assessment* of creativity may not be necessary (or practical) in many higher-education learning contexts. However, engaging students in assessing their own creativity in ways that lead to informed recognition of their creative effort, in order to produce useful *formative* feedback for the learner and the teacher, should be an essential feature of an assessment system to promote students' creativity and the development of self-awareness of their own creativity. This view mirrors that developed independently by John Cowan in Chapter 12 of this book. The process of students getting together to discuss their productions or creative processes, and co-creating criteria within a framework of prompts to guide conversation, would not only allow them to reflect on their own creative effort, but would enable them to construct new understandings of what creativity means in the particular contexts in which it was utilised. It would also enable them to represent their own creativity and, ultimately, to harness their own creativity to a higher degree. Consensual assessment involving teachers and students could be integrated into the sort of assessment process described by John Cowan.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46