The longer term value of creativity judgements in computational creativity

Anna Jordanous [1]

Abstract. During research to develop the Standardised Procedure for Evaluating Creative Systems (SPECS) methodology for evaluating the creativity of 'creative' systems, in 2011 an evaluation case study was carried out. The case study investigated how we can make a 'snapshot' decision, in a short space of time, on the creativity of systems in various domains. The systems to be evaluated were presented at the International Computational Creativity Conference in 2011. Evaluation was performed by people whose domain expertise ranges from expert to novice, depending on the system. The SPECS methodology was used for evaluation, and was compared to two other creativity evaluation methods (Ritchie's criteria and Colton's Creative Tripod) and to results from surveying people's opinion on the creativity of the systems under investigation. Here, we revisit those results, considering them in the context of what these systems have contributed to computational creativity development. Five years on, we now have data on how influential these systems were within computational creativity, and to what extent the work in these systems has influenced further developments in computational creativity research. This paper investigates whether the evaluations of creativity of these systems have been helpful in predicting which systems will be more influential in computational creativity (as measured by paper citations and further development within later computational systems). While a direct correlation between evaluative results and longer term impact is not discovered (and perhaps too simplistic an aim, given the factors at play in determining research impact), some interesting alignments are noted between the 2011 results and the impact of papers five years on.

1 Introduction

In [8], the Standardised Procedure for Evaluating Creative Systems (SPECS) methodology was developed as a tool for evaluating the creativity of software developed within computational creativity research. SPECS is summarised in Table 1. As part of the research to develop SPECS, two case studies were carried out; this paper focuses on the second case study reported in [8].

The case study we focus on here was carried out at the 2011 International Computational Creativity Conference (ICCC'11), and explored to what extent creativity evaluation methods can be applied across creative systems demonstrating different types of creativity rather than focusing exclusively on systems operating specifically within one creative domain. This case study specifically explored the scenarios where we do not have the full information desired for evaluation, or where we may have only limited time to complete evaluation, or be limited as to who can perform evaluation. This was motivated by the issue that we often wish to compare one system's creativity against that of others, but for various reasons may not have the full information available to us that we would like, or may be working under time pressures. Section 1.1 discusses this.

Four different evaluation methods were applied to evaluate the creativity of five systems: the collage generation module for the artistic system *The Painting Fool* [2] [4] [4]; a poetry generator [17]; the *DARCI* system [15] for generating images to illustrate given adjectives; a reconstruction of the *MINSTREL* story-telling system [23] [22]; and a musical soundtrack generator matching emotions in a narrative to the music generated [12]. The evaluation methods used in this case study were: SPECS [8, 7]; Ritchie's empirical criteria [18]; Colton's creative tripod [2]; and asking people's opinion on how creative each system was. In each evaluation, the judges worked with limited information and time.

The 2011 case study resulted in formative evaluative feedback for the systems to help researchers develop the creativity of their system. Section 2 summarises the generated feedback, which is fairly detailed even given the time and information pressures.[2] As the creative domain varies across systems, comparisons between systems became less relevant: the systems were designed to perform different tasks, requiring different interpretations of creativity [16, 8]. Hence the focus in this case study was on evaluating individual systems - though some interesting comparisons could made between systems where there are commonalities in creative priorities of that domain.

Looking back at this case study five years later, we can now see what contributions each system has made to the development of computational creativity research over the past five years, as measured in citations each 2011 paper has received, and in tracing what development work has been done since 2011 that can be directly related to the 2011 systems. This is a beneficial exercise: given that value is an important part of creativity, we could hypothesise that those systems judged more creative can have had more value to the computational creativity community over the last few years. Hence we can test on our (small)[3] sample as to whether initial judgements of the creativity of these systems give us information as to which systems will provide greater contributions to computational creativity research.

1.1 Digital resource availability for evaluation

Creative systems are by their nature likely to be different to every other system and it is useful to see how a creative task has been approached in different ways, when we are investigating that task

[1] School of Computing, University of Kent, Medway Building, Chatham Maritime, Kent, UK, email: a.k.jordanous@kent.ac.uk

[2] An unexpected but beneficial extra finding of the evaluation was that it highlighted which ICCC'11 presentations had contained adequate information for judging the creativity of their systems.

[3] Sample size means that results are indicative rather than conclusive.

Table 1. The Standardised Procedure for Evaluating Creative Systems (in summarised form)

1. Identify a definition of creativity that your system should satisfy to be considered creative
2. Using Step 1, clearly state what standards you use to evaluate the creativity of your system.
3. Test your creative system against the standards stated in Step 2 and report the results.

comptationally. There may be systems that are related in some way to systems that we are developing, where it would be of interest to learn more about the research behind that system(s); in particular it would be useful to gain knowledge from seeing the system in operation, as well as reading published reports. As example, in evaluating the GAmprovising musical improvisation system against GenJam and Voyager [8], several useful insights arose for the development of GAmprovising from evaluating GenJam and Voyager.

It is more straightforward to evaluate systems for which we have full access to view and run the source code, with as much time available as we need, all necessary data and a line of communication with the system developers. This ideal evaluation scenario, however, is often not possible.

Taking time constraints first, the amount of time researchers can spend on evaluation is partly dictated by factors such as the allocation of researchers' time (particularly when conducting multiple projects or when time allocations are dictated by funding), deadlines for conferences etc., time demands within a project and the scheduling of other tasks to be completed within the project. Further demands on researchers' time include teaching, administration and other research work. There are often also constraints on the time and availability at appropriate times of other people involved in performing the evaluation. Another important issue impacting upon evaluation is if there are problems with availability of relevant software, data or more detailed information for a creative system(s) that we are interested in.

We could choose not to use systems for comparative evaluation if we do not have the full access and data that we would like; however, while this reduces the evaluation workload, it also sacrifices the opportunity to learn from this system. Alternatively, we can include systems in comparative evaluation even if we only have partial information for that system, keeping aware of the constraints on what we can learn from such evaluation but taking advantage of what is available, for formative feedback into the development of our own existing and future systems. Without evaluation of other systems:

> 'lessons from the past are difficult to learn' [1, p. 149, reflecting on the lack of availability of computer artworks and their related research resources]
> 'without cultural artifacts, civilization has no memory and no mechanism to learn from its successes and failures. And paradoxically, with the explosion of the internet, we live in what Danny Hillis has referred to as our "digital dark age".'[4]

When would we wish to learn from other existing systems? Systems of historical interest would have intrinsic value, even if the system can no longer be obtained. For example, James Meehan's TALESPIN system [11] has proven to be a seminal work in the field of story generation, even though the code has been lost and only a 'micro-TALESPIN' version exists today [10], which was itself published over 30 years ago. Similarly, Christopher Longuet-Higgins produced software for expressive musical performance which was widely praised by those who heard it [5, and personal communications with: Darwin, 2012; Dienes, 2012; Torrance, 2012; Thornton, 2012]. Unfortunately, the system was not made available as code or in a published report before Longuet-Higgins' death, and the code was archived but now cannot be restored due to the use of obsolete data storage formats.[5] We can learn from what our peers are doing in closely related research areas, and also by cross-applying work from less related areas to our own interests.

As Robey [19] has remarked, research that produces computer programs is surrounded by issues of software sustainability. Unfortunately, even for more recent systems, it can be difficult to retrieve all information necessary for full evaluation of a system. Bentkowska-Kafel [1] and Robey [19] have highlighted the speed at which current or cutting-edge digital resources can quickly become obsolete or lost, sometimes in a matter of only a few years.

> 'digital information lasts forever - or five years, whichever comes first.' [20, p. 2]

Jordanous [8] discusses several potential reasons:[6]

- Digital resources such as source code may not have been made available publicly.
- The researchers may not be available to contact (they may have left academia, or passed away) or may have moved onto other projects and forgotten details of the project of evaluative interest.
- Code may be unavailable or obsolete even if obtained. [6, pp. 34-35] identifies various reasons why available code may become unusable, including hardware or software obsolescence, third party dependencies, proprietary or poorly documented code as well as concerns about protecting intellectual property rights (especially in more competitive scenarios).
- Published code/digital resorces may not remain available long-term, for example if funding runs out for online hosting costs.

1.2 Selection of the creative systems being evaluated

The International Computational Creativity Conference (ICCC) is an annual international conference series dedicated to computational creativity research. Since its inception in 2010 it has been the main presentation venue for the latest findings in computational creativity research, taking over this role from the previous International Joint Workshops in Computational Creativity (IJWCC), from which the conference series evolved. ICCC'11 was held in Mexico in April 2011. Many creative systems were presented, demonstrating various types of creativity in different domains.

At ICCC'11, papers were presented to the conference audience in talks lasting seven minutes (a particularly brief amount of time for talks). There is a limit to what can be presented in this time and

[4] Original source unattributed, quote taken from $http://archive.org/about$ (last accessed January 2016).

[5] According to personal communications with Jeremy Maris and other IT support staff at the University of Sussex, where Longuet-Higgins' computer files were archived, and with a digital archive specialist, Gareth Knight.

[6] Jordanous [8] also discusses several recent efforts to promote software sustainability.

Table 2. The five systems from ICCC'11 that were evaluated for the 2011 case study.

Paper	System (if named)	Domain	Purpose
[4]	Module of *The Painting Fool*	Art	Collage generation
[17]	Adapted from an earlier system: McGonagall [9]	Poetry	Poetry generation
[15]	*DARCI*	Art	Image selection
[22]	Reconstruction of *MINSTREL*	Narrative	Story generation
[12]	-	Music	Soundtrack generation

it is unlikely that all desired information can be provided, but the ICCC'11 organisers posited that enough information could be delivered for the audience to become acquainted with the paper content. During the seven-minute talks at ICCC'11, presenters aimed to convey enough information about their paper so that people could discuss issues raised, in a group of related talks.

Five of the creative systems presented at ICCC'11 were selected for the 2011 case study, representing a variety of creative domains. These five systems were evaluated by two judges on their creativity using SPECS [8, 7, See Table 1 for a summary], based on the information presented in the seven-minute talks. The evaluation also generated qualitative feedback for the presenters of the evaluated systems, from two perspectives: the perceived creativeness of their system and the quality of information that they presented about their system in the brief time frame permitted. (The original purpose of this case study in 2011 was to test out the SPECS methodology for evaluating creative systems.)

One more point to note is that the systems evaluated in this case study were from a variety of domains, rather than just one domain. Some comparisons can be made between systems from different domains, and some interesting insight can be gained from doing so. On the whole, though, this paper acknowledges that such comparisons are less useful than comparisons made between systems from similar domains, as there are fewer areas of crossover so therefore less relevant information is available from the comparison. Some non-obvious conclusions may still however be reached this way, through viewing the systems from different perspectives. Comparing systems across different creative domains can also give a general (if limited) impression of relative progress in each domain.

Later, the information collected in SPECS evaluation was re-applied to use Colton's Creative Tripod as the evaluative method. Then analysis was carried out on the five selected systems using Ritchie's empirical criteria [18], and through asking people their opinion on systems' creativity. To replicate the time pressures and limits on available information in the latter stage of evaluation, judges were given only the information available in the paper, and had only a short amount of time to read the paper before evaluation.[7][8]

Of the five presentations in the first session (entitled 'The Applied'), the judges decided that three presented details of a computational creativity system that could be evaluated. Two further systems from the third session ('The Narrative') were also evaluated, for a total of five systems evaluated in this case study. These five systems, along with the papers they were presented in, the authors and the creative domain which the system operates in, are listed in Table 2.

[7] It is acknowledged that a closer replication could have been achieved; however the principles behind the evaluation remain the same (time and information availability pressures) and in this paper the emphasis is not on comparing the different evaluation methods, but on learning from their collective findings.

[8] Full details of the methodologies and how they were applied in this case study can be found in [8].

Examples of some of these systems' output are given in Figures 1 [4] and 2(a) [15], and in this example limerick poem from [17]:

'There **was** the young **la**dy called **Bright**.
They could **tra**vel much **fas**ter than **light**.
They **re**latively **left** one **day**. It sur**vived**.
They **left**. On the **night**, she re**turned**.'
[17, p. 9] *Where syllables are in bold, that syllable should be stressed when reading aloud the limerick.*

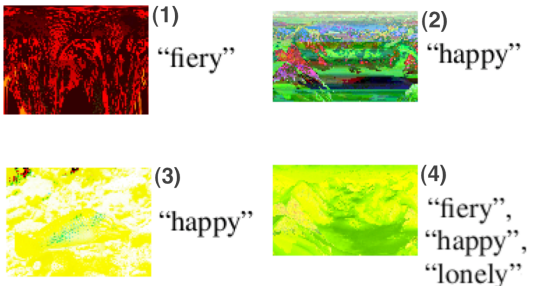
Figure 1. Collage generated by Cook & Colton's collage generation module in *The Painting Fool* system, on the theme of the current war in Afghanistan (Cook & Colton, 2011, Fig. 1, p. 2).

2 Results of the 2011 case study

2.1 Applying the SPECS methodology

The 14 components of creativity identified in [7] (see Figure 3) were used as a definition of creativity in a general context.

The judges recorded what general creative domain each system was designed to operate in (e.g. art, narrative generation). They also assessed their level of expertise and competence in that domain as either *Basic*, *Reasonable* or *Expert*. For each component, judges rated how successfully the system performed on that component requirements, giving a score between 0 (lowest) and 10 (highest). The judge rated the system based on the information given in the conference



(a) Output images, each intended to illustrate the adjective(s) listed to the right of that image (Norton et. al., 2011, Figs. 4-7, p. 14).

(a) Image A (b) Image B (c) Image C

(b) Inspiring set images (Norton et. al., 2011, Fig. 2, p. 13).

Figure 2. *DARCI* output, and the inspiring set of source images used to generate this output.

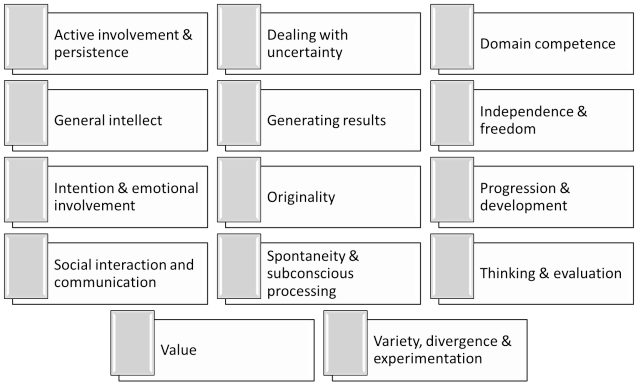


Figure 3. Jordanous's 14 components of creativity [7], derived through empirical analysis of the words used in texts about creativity.

talk; if they felt that not enough information was given about a particular component to provide a rating, then this rating was left blank. Each component was categorised according to how important the judge felt that component was for creativity in the domain which that system operated in. The contribution of that component to creativity in the system's domain was categorised according to how important that component was for creativity.

Jordanous [8] presents full results of what was learned from this case study; here our primary focus is in seeing how the creativity of each system was judged, relative to the other systems. To summarise:

- The collage generator [4] performed well at creating results, demonstration of intention and social abilities, but could improve its originality, value and ability to be spontaneous.
- The poetry generator [17] was good at creating results in a domain-competent way but needs to improve its ability to experiment and diverge and to a lesser extent could improve upon its originality, value and spontaneity.
- *DARCI* [15] showed strengths in social interaction, spontaneity, self-evaluation and production of results, but could perform better on originality and value.
- The story generator's [22] abilities to be original and to produce results were praised, though it could improve upon its inherent value, its ability to progress and develop and to work independently.
- The soundtrack generator [12] was considered valuable and competent in its domain, but could improve its ability to diverge and experiment.

Some systems performed better in evaluation, notably *DARCI*:

- *DARCI* [15] was rated highly on 50% of the components key to creativity in its domain, with the remaining systems scoring between 25% [4] and 33% [17, 12].
- Accounting for middling ratings as well, again *DARCI* was ahead of the other systems, with 75% of its key components receiving a high or middling rating. Three systems had 50-54% of its key components receiving high or middling ratings [17, 22, 12]. *The Painting Fool*'s collage generator [4] only received high or middling ratings for 25% of its key components.
- The reconstruction of *MINSTREL* [22] was the only system to receive a low rating for one of its key components, though it did have the largest number of key components to address.
- Quantifying the ranking data obtained such that high ratings score 2 points, middling ratings score 1 point and low ratings or unrated components score 0 points, with the total divided by the number of components considered key to that type of creativity by the judges,[9] overall rankings can be generated:

1. *DARCI*: 5/4 = 1.25 points.
2. *MINSTREL* reconstruction: 6/7 = 0.857 points (to 3 s.f.).
3. by Rahman & Manurung: 5/6 = 0.833 points (to 3 s.f.) and by Monteith et al.: 5/6 = 0.833 points (to 3 s.f.).
4. by Cook & Colton: 2/4 = 0.5 points.

2.2 Applying Ritchie's criteria

Ritchie's criteria were applied in a similar fashion to the applications reported in [18], except that (because of the Boolean way they are defined by Ritchie) the criteria were treated as a set of criteria which can be either satisfied or not satisfied, depending on whether a threshold value is reached or not. This approach better fits the As discussed in [8], Ritchie's criteria [18] concentrate almost exclusivelyon observations about the output of the system, measuring how typical that output is of the domain and how valuable the output is in the domain. (The criteria also include information on the *inspiring set* of input examples a system may have been constructed from.) An approach similar to that used in SPECS was adopted to meet these demands, with two judges asked to provide ratings. If the authors of the five of the 2011 case study systems had all provided examples of their systems (or links to examples) in their papers, then these could be used for evaluation using Ritchie's criteria. Ideally, details of inspiring sets would also be available for Ritchie's criteria to be fully applied. Unfortunately this was not always the case. Jordanous [8] discusses reasons for variability in available information, and reports efforts to locate additional examples of output and inspiring sets. It is

[9] This is of course one of several ways to quantify this information.

worth repeating here, however, that the point of this evaluation case study was to evaluate systems based on the information available, and work with incomplete information.

What can be done is to evaluate the results of the systems as and when presented in the papers, with no evaluations being performed for the Tearse et al. [22] and Monteith et al. [12] systems using Ritchie's criteria as these authors did not provide examples either in their papers or in online supplementary resources. Full details of the criteria calculations for each system are given in [8]. To summarise:

- For the collage generator, of 8 applicable criteria, 1 evaluated as TRUE (Criterion 10a) and 7 as FALSE (Criteria 1-4, 6, 7, 9).
- For the poetry generator, of 7 applicable criteria, 1 evaluated as TRUE (Criterion 10a) and 6 as FALSE (Criteria 1-4, 6, 9).
- For *DARCI*, of 10 applicable criteria, 2 evaluated as TRUE (Criteria 5, 10a) and 8 as FALSE (Criteria 1-4, 6-9).
- Neither the story generator or the soundtrack generator could be evaluated due to lack of information on their inspiring sets.

DARCI [15] was the only system for which two criteria (5, 10a) rather than one (10a) were true. It also had the fewest inapplicable criteria; the only inapplicable criteria were Criteria 11-18 which, it was noted earlier, could not be applied for any of these systems if the results set did not include items from the inspiring set.

The two criteria that *DARCI* satisfied were:

5. A decent proportion of the output are both suitably typical and highly valued.

10a. Much of the output of the system is not in the inspiring set, so is novel to the system.

The two other evaluated systems [4, 17] also satisfied the second of these criteria, 10a.

It is unclear in [18] how the criteria results should be analysed. Is *DARCI* [15] the most creative because it satisfied 2 criteria as opposed to 1, or is Rahman & Manurung's poetry generator [17] poetry generator most creative because it had least false criteria (6 as opposed to 7 for Cook & Colton [4] and 8 for Norton et al. [15])? Or should the number of inapplicable criteria be taken into account? It was decided that for this analysis, the percentage of applicable criteria that were true would be calculated for each system and this would be used to compare the systems' creativity. Therefore if a criterion is not applicable to a system, it is not considered for that system.

- Cook & Colton's collage generator [4] satisfied 1 out of 8 applicable criteria (12.5%).
- Rahman & Manurung's poetry generator [17] satisfied 1 out of 7 applicable criteria (14.3%).
- Norton et al.'s image generator [15] satisfied 2 out of 10 applicable criteria (20%).

These results place the *DARCI* image generator by Norton et al. [15] as the most creative system of the three, followed by Rahman & Manurung's poetry generator [17] and then Cook & Colton's collage generation module for *The Painting Fool* system [4]. For all three systems, though, only a small percentage of criteria were satisfied.

2.3 Applying Colton's creative tripod

In applying the creative tripod [2] for evaluation, we see that Colton's tripod qualities map to three of the 14 components used for SPECS:

- Skill $\approx$ *Domain Competence*.
- Imagination $\approx$ *Variety, Divergence and Experimentation*.
- Appreciation $\approx$ *Thinking and Evaluation*.

The evaluation data gathered on these three components could therefore be used to evaluate the systems using the creative tripod.

- The collage generator showed average imaginative abilities and there was a lack of information on other qualities, with mean ratings out of 10 of 5.0 for imagination but no data for skill or appreciation.
- The poetry generator demonstrated very good skilfulness and appreciation with average imagination, with mean ratings out of 10 of 8.5 for skill, 5.0 for imagination and 8.0 for appreciation.
- *DARCI* showed average to good all-round performance on the tripod qualities, with mean ratings out of 10 of 7.0 for skill, 7.5 for imagination and 6.0 for appreciation.
- The story generator performed reasonably well on all three tripod qualities although could improve its imaginative abilities, with mean ratings out of 10 of 8.0 for skill, 5.5 for imagination and 7.0 for appreciation.
- The soundtrack generator gave partial information on the tripod qualities, demonstrating average skill and imagination, with mean ratings out of 10 of 6.0 for skill, 5.0 for imagination but no data for appreciation.

Three systems emerge from this evaluation as 'balanced', i.e. with all three 'legs' present [17, 15, 22]. Monteith et al.'s system [12] could not be evaluated on its appreciative abilities and Cook & Colton's system [4] presentation lacked information on both its skill and appreciation. Both systems received only middling ratings in all cases where ratings were provided, apart from a 7/10 for Monteith et al.'s system's [12] skill from one judge (accompanied by a 5/10 from the other judge).

Taking the mean of the three qualities for each system, overall averages were 7.2 for Rahman & Manurung [17] and 6.8 each for Norton et al. [15] and Tearse et al. [22]. These observations indicate that Rahman & Manurung's system [17] was found to be more creative than the other systems, as it had a higher mean overall and the highest ratings for two out of three qualities. Its performance for appreciation, though, was only average (5/10). It could be argued that *DARCI* demonstrated a better all-round performance and was therefore found to be more creative.[10] The other two systems [4, 12] were considered less creative overall than these three systems, because they did not demonstrate clear abilities on some of the tripod qualities (given the information in the presentations on the systems). Of these two, Monteith et al.'s system [12] may have been slightly superior to that of Cook & Colton [4] because it demonstrated some aspects of both skill and imagination and received one rating of 7/10 (from Judge 2 for skill) in comparison to the rest of the ratings for these two systems (either left blank or rated as 5/10).

We can conclude that with the above use of the Creative Tripod, Rahman and Manurung's system performed best in creativity evaluation, followed jointly by Norton et al. [15] and Tearse et al. [22], then Monteith et al.'s system [12], and finally Cook & Colton's system [4].

[10] Given that Colton [2] does not investigate how to use the creative components for quantitative comparison, and that no such usage of the creative tripod was found to use as an example, exact conclusions are speculative. it is noted here that Colton's Creative Tripod is intended to identify whether computational systems can be considered candidates for potentially creative systems, rather than evaluating their creativity per se, so this case study stretches the application of the Creative Tripod somewhat beyond what Colton originally intended.

2.4 Collecting people's opinions of creativity

The evaluation results and feedback obtained for the 2011 case study were compared to human evaluations of the creativity of the case studies. The two judges were asked to say how creative each system was and their reasons and comments. Judges could choose from the following options to describe a system's creativity: Completely creative; Very creative; Quite creative; A little creative but not very, Not at all creative. More relevant to this paper's investigations, judges were also asked to rank the five systems in terms of relative creativity.

- The collage generator: 'Not at all creative/Quite creative'. The complexity of the processes used was praised by one judge but seen as trivial for creativity by the other.
- The poetry generator: 'A little creative but not very/A little creative but not very'. It generated interesting poetry but did not generate what was intended and was more aimed at generating a target example.
- *DARCI*: 'A little creative but not very/Quite creative'. *DARCI*'s ability to learn was highlighted as a useful attribute by one judge. The system may be more useful for interactive creativity with humans than as a standalone system, though, with one judge seeing the creativity of DARCI as located within the knowledge obtained from people.
- The story generator: 'A little creative but not very/Quite creative'. Whilst creating stories that seem to be fairly interesting but slightly nonsensical, the process did not seem to be optimised for increasing the interestingness of its stories, but for replicating as closely as possible a previous system (MINSTREL).
- The soundtrack generator: 'Quite creative/Quite creative'. Judges liked the intentions behind the system and its ability to combine two existing systems and layer human involvement, but needed more information for a fuller opinion.

An overall ranking of systems' creativity can be generated from the data on the judges' rankings, in Table 3. For each judge's rankings, 5 points were allocated to the system ranked most creative, down to 1 point for being ranked least creative. The two sets of ranking points were then summed together:

- Monteith et al. [12] was ranked most creative overall with 10 points (5+5).
- Norton et al. [15] was ranked second most creative overall with 7 points (3+4).
- Tearse et al. [22] was ranked third most creative overall with 5 points (3+2).
- Rahman & Manurung [17] and Cook & Colton [4] were ranked joint least creative overall with 4 points (3+1 and 1+3 respectively).

The ratings and feedback show some common opinions between the judges. For example, both judges praised the processes involved in Monteith et al. [12] and both criticised Rahman & Manurung [17] for their focus on replicating a target poem rather than creating new poetry. Individually, judges' opinions varied a great deal, as is perhaps to be expected with using only two judges who have differing backgrounds and expertise in the various creative domains covered.

One thing this has illustrated is that individual people can form very different first impressions of systems. Taking two people's opinions was useful for directed, constructive criticism, but was less useful for any significant statements about which systems are more or less creative than each other, compared to the more formal evaluation methodologies employed. Though a similar previous case study with 111 recipients [8] has suggested a larger number of judges does not necessarily produce conclusive distinctions between systems' perceived creativity, this evaluation for the 2011 case study showed the limits of what can be taken from a small number of judges.

Table 3. Ordering the case study systems by creativity: Judges' opinions.

Creativity	Judge 1	Judge 2
Most: 1	Monteith et al. [12]	Monteith et al. [12]
2	{ Rahman & Manurung [17] / { Norton et al. [15] / { Tearse et al. [22] (equal)	Norton et al. [15]
3	"	Cook & Colton [4]
4	"	Tearse et al. [22]
Least: 5	Cook & Colton [4]	Rahman & Manurung [17]

3 Comparing the success of different evaluation methods in the two case studies

Four different evaluation methods have been now used to evaluate the creativity of the five systems in the 2011 case study: SPECS (using the components in [7] as a basic definition of creativity); people's opinions of creativity of a system; Ritchie's empirical criteria and Colton's creative tripod. Each generated evaluative information on each system and comparisons of creativity between systems within each case study.[11] Here we focus on the relative rankings the evaluation methods generated for the five evaluated systems

3.1 Comparing evaluation results and feedback in the 2011 case study

No two methodologies produced the same results, but typically, *DARCI* [15] was judged one of the more creative systems and the collage generation module for *The Painting Fool* was judged one of the less creative systems. Otherwise, there was some disagreement between findings, largely caused by the lack of a 'ground truth' or baseline to judge the systems and the different domains that the systems work in. The priority in this case study in 2011 was on obtaining formative feedback for the system authors.

Here, we look at the results of each evaluation method to see if they can help us predict which systems have had a longer term contribution to the field of computational creativity. This is somewhat akin to the way in which we judge systems based on their presentation at a conference. While there are many factors beyond a conference presentation that determine what weight we give to a work's contribution, a major resource for information in computational creativity research comes from the information available at the key international conference for this research, the ICCC conference series. Both the papers and the presentations form key sources of information for computational creativity researchers; this is also true for the information in the 2011 case study.

Two ways in which we can measure contribution of papers to computational creativity research are (1) to count a paper's citations; and (2) to examine citations to see if those systems in 2011 have had direct influence in further computational creativity research.

[11] A full presentation and discussion of this information can be found in [8].

Table 4. Overall comparisons of the relative creativity of each system in the case study from 1: Most creative to 5: Least Creative. NB 'gen': 'generator'.

Evaluation Method	SPECS using components	Survey of opinions	Ritchie's criteria	Colton's tripod
1	*DARCI*	soundtrack gen	*DARCI*	poetry gen
2	story gen	*DARCI*	poetry gen	*DARCI* / story gen
3	poetry gen / soundtrack gen	story gen	collage gen	*DARCI* / story gen
4	poetry gen / soundtrack gen	poetry gen / collage gen	- (other two systems unrated)	soundtrack gen
5	collage gen	poetry gen / collage gen	- (other two systems unrated)	collage gen

3.2 Citation counts

Using Google Scholar, we can see how many citations the papers reporting each system under investigation have attracted since 2011 (reported in descending order of total citation count): see Table 5.

Table 5. Number of citations for each paper in the 2011 case study (in descending order of total citations, according to Google Scholar):

Paper	# citations	# non-self-citations
Norton et al. [15]	13	5 (38%)
Cook & Colton [4]	11	3 (27%)
Rahman & Manurung [17]	8	8 (100%)
Tearse et al. [22]	7	3 (43%)
Monteith et al. [12]	5	4 (80%)

Table 5 shows that Norton et al.'s work [15] has received the most citations overall, followed by Cook & Colton's work [4]. Monteith et al.'s paper [12] has to date received the fewest citations. If we look at non-self-citations, i.e. those citations from papers with no shared authors to the original paper, then Rahman & Manurung's work emerges as highest in influence both in terms of number of non-self-citations and overall percentage of non-self-citations compared total citations. At the other end of the scale, Cook & Colton's paper and Tearse et al.'s paper both attract only 3 non-self-citations.

We should remember the number of factors involved in citations: such as is the citation positive or negative? does the citation focus on the work cited or merely acknowledge it in passing? how active are the original authors in publishing their own work at other venues? But nonetheless, citation counts continue to be a key metric in measuring research impact. The absolute number of citations highlighted DARCI [15] (in rough alignment with the evaluation methods) and Cook & Colton's Painting Fool module [4] (not in alignment with the evaluation methods. Perhaps more importantly for this metric, the number of non-self citations highlighted Rahman & Manurung's poetry generator (in rough alignment with all the formal evaluation methods, though not the opinion-based evaluation), with Norton et al.'s paper on DARCI receiving the second highest number of non-self-citations (roughly aligning with all results from the 2011 study).

3.3 Direct influence on subsequent computational creativity research

What current (or successive) work did the 2011 papers inform? This is where citation data from both self-citations and non-self-citations can be investigated more thoroughly. We find that (in rough order of the 2011 case study rankings, across all evaluation methods):

- DARCI [15] is still being featured in subsequent publications in most years, with an active online community crowdsourcing data for DARCI's development. [15] is also cited as influencing work on other systems [14, 21, for example].
- Rahman & Manurung's poetry generator [17] has been cited across papers reporting multiple different pieces of work [13, 3, 21, for example]. The first author of this work has not since published work in the computational creativity field, unlike the other authors, but the work has clearly made some impact on the computational creativity field. The second author has since published work in computational creativity, but interestingly, has not since cited this 2011 paper.
- Tearse et al.'s story generator [22] has mostly been cited in papers considering further development of the MINSTREL reconstruction: showing influence in creativity development through one system, but not a wider impact (to date).
- Monteith et al.'s soundtrack generator [12] has been cited in reports of other systems, with some influence evident in the way the system is reported in some of these citations. The paper has not, however, been cited by the authors themselves, suggesting that development of this particular system has taken different paths since 2011.[12]
- Cook & Colton's collage generator [4] is arguably part of one of the most prominent systems longer term, being a module for the Painting Fool artistic system. This system has attracted much publicity through exhibitions, further publications, and public engagement/impact activities, though it is unclear whether the collage generator module is influencing this system, or whether it is a module that may or may not be used depending on the current application of the Painting Fool.

What we see here is that the DARCI system has again been recognised as valuable computational creativity software. Rahman and Manurung's poetry generator has also been found to have longer-lasting influence across computational creativity work, even though the lead author of this paper is not a regular participant in computational creativity research events.

The 'surprise' result given the 2011 case study results (when considered in isolation) is the long-lasting impact of Cook & Colton's collage generator. This reminds us that it is not merely an evaluation

[12] Perhaps somewhat ironically, two of the authors do however cite Rahman and Manurung's work in a later paper of theirs.

of a system which can help us judge the longer-term impact of a creative system in computational creativity research; there are several other factors in play such as the provenance of the system (e.g. its authors, what system(s) it is derived from). It is interesting, however, to see some consistency in alignments between the 2011 evaluations and the metrics employed here for longer term impact.

4 Summary

The 2011 case study carried out during the development of the Standard Procedure for Evaluating Creative Systems (SPECS), showed how various computational creativity evaluation methods could be applied to evaluate the creativity of various types of creative systems from different creative domains. This 2011 case study captured first impressions and initial evaluations of how creative systems were, with limited information and resources, and under time pressures.

Analysis of these evaluations provided information about how creative the systems were perceived to be and what information contributed to this, relative to the creative domain. This 2011 case study also highlighted what information is most useful to help people make evaluations of creativity based on conference papers and presentations - key sources of information for computational creativity researchers. Several evaluation methods were applied to the systems evaluated in the 2011 case study. As well as SPECS [7, 8], people's opinions were consulted on the creativity of the systems. Two key existing methodologies for computational creativity were also applied: [18, 2, Ritchie's criteria and Colton's creative tripod, respectively]. Results were compared; it was noted that few 'right answers' or 'ground truths' for creativity were found in the 2011 case study.

The consequences of judging a system given limited and perhaps incomplete information meant that occasionally important information for evaluation is missing. This affected the use of all the evaluation strategies employed. It is interesting to see which methodologies were most robust when dealing with missing information. Collecting people's opinions seemed the best approach at coping with missing information, as might be expected given that little was specified for the humans to use as a definition of creativity. SPECS was relatively robust, as was Colton's tripod framework. Ritchie's criteria approach was the most affected by missing information, as various criteria could not be applied and the absence of information on inspiring sets and example outputs had significant effects.

Looking longer term at whether initial evaluations of creativity can help us predict which systems are likely to make a longer term contribution to creativity: this has been an interesting experiment. Certainly, some alignment was found between the 2011 evaluation results and the impact five years on that each system/paper has made, as can be measured by citation quantity and types. However the evaluation results did not directly correlate with study of later impact. The creativity of systems presented in computational creativity is one factor which contributes to their value to the community, but as discussed above, it is of course not the only factor. However, evaluation methods are giving us some hints for gauging longer term impact. This experiment has only looked at impact over a five year period. Perhaps in ten years (at AISB'21?), or over even longer time periods, we will uncover different results?

REFERENCES

[1] Anna Bentkowska-Kafel, 'The Fix vs. the Flux: Which digital heritage?', in *Netpioneers 1.0 - archiving, representing and contextualising early netbased art*, eds., Dieter Daniels and Günther Reisinger, 145–162, Sternberg Press in association with the Ludwig Boltzmann Institute, Berlin, Germany / New York, NY, (2009).

[2] Simon Colton, 'Creativity versus the Perception of Creativity in Computational Systems', in *Proceedings of AAAI Symposium on Creative Systems*, pp. 14–20, (2008).

[3] Simon Colton, Jacob Goodwin, and Tony Veale, 'Full-FACE poetry generation', in *Proceedings of the International Conference on Computational Creativity*, pp. 95–102, Dublin, Ireland, (2012).

[4] Michael Cook and Simon Colton, 'Automated Collage Generation - With More Intent', in *Proceedings of the 2nd International Conference on Computational Creativity*, pp. 1–3, Mexico City, Mexico, (2011).

[5] Chris Darwin. Obituary: Christopher Longuet-Higgins. *The Guardian*, June 10th, 2004, jun 2004.

[6] Neil Chue Hong, Steve Crouch, Simon Hettrick, Tim Parkinson, and Matt Shreeve, 'Software Preservation: Benefits framework', Technical report, Software Sustainability Institute and Curtis & Cartwright Consulting Ltd., (dec 2010).

[7] Anna Jordanous, 'A Standardised Procedure for Evaluating Creative Systems: Computational Creativity Evaluation Based on What it is to be Creative', *Cognitive Computation*, **4**(3), 246–279, (2012).

[8] Anna Jordanous, *Evaluating Computational Creativity: A Standardised Procedure for Evaluating Creative Systems and its Application*, Ph.D. dissertation, University of Sussex, Brighton, UK, sep 2012.

[9] Hisar Maruli Manurung, *An Evolutionary Algorithm Approach to Poetry Generation*, Ph.D. dissertation, School of Informatics, University of Edinburgh, Edinburgh, UK, 2003.

[10] James Meehan, 'Tale-Spin', in *Inside computer understanding: five programs plus minatures*, eds., R C Schank and C K Riesbeck, Lawrence Erlbaum Associates, Hillside, NJ, (1981).

[11] James Richard Meehan, *The metanovel: writing stories by computer.*, Ph.D. dissertation, Yale University, New Haven, CT, USA, 1976.

[12] Kristine Monteith, Virginia Francisco, Tony Martinez, Pablo Gervás, and Dan Ventura, 'Automatic Generation of Emotionally-Targeted Soundtracks', in *Proceedings of the 2nd International Conference on Computational Creativity*, pp. 60–62, Mexico City, Mexico, (2011).

[13] Kristine Monteith and Tony Martinez, 'Automatic generation of melodic accompaniments for lyrics', in *Proceedings of the 3rd International Conference on Computational Creativity*, Dublin, Ireland, (2012).

[14] R G Morris, S H Burton, P M Bodily, and D Ventura, 'Soup Over Bean of Pure Joy: Culinary Ruminations of an Artificial Chef', in *International Conference on Computational Creativity*, p. 119, (2012).

[15] David Norton, Derrall Heath, and Dan Ventura, 'Autonomously Creating Quality Images', in *Proceedings of the 2nd International Conference on Computational Creativity*, pp. 10–15, Mexico City, Mexico, (2011).

[16] Jonathan A Plucker and Ronald A Beghetto, 'Why Creativity is Domain General, Why it Looks Domain Specific, and why the Distinction Doesn't Matter', in *Creativity: From Potential to Realization*, eds., Robert J Sternberg, Elena L Grigorenko, and Jerome L Singer, chapter 9, 153–167, American Psychological Association, Washington, DC, (2004).

[17] Fahrurrozi Rahman and Ruli Manurung, 'Multiobjective Optimization for Meaningful Metrical Poetry', in *Proceedings of the 2nd International Conference on Computational Creativity*, pp. 4–9, Mexico City, Mexico, (2011).

[18] Graeme Ritchie, 'Some Empirical Criteria for Attributing Creativity to a Computer Program', *Minds and Machines*, **17**, 67–99, (2007).

[19] David Robey. Introduction to Digital Humanities. Talk, Sep 2011.

[20] Jeff Rothenburg. Ensuring the Longevity of Digital Information. Available at http:// www.clir.org / pubs / archives / ensuring.pdf (last accessed November 2012), feb 1999.

[21] M. R. Smith, R. S. Hintze, and Dan Ventura, 'Nehovah: A neologism creator nomen ipsum', in *Proceedings of the 5th International Conference on Computational Creativity*, Ljubljana, Slovenia, (2014).

[22] Brandon Tearse, Peter Mawhorter, Michael Mateas, and Noah Wardrip-Fonin, 'Experimental Results from a Rational Reconstruction of MINSTREL', in *Proceedings of the 2nd International Conference on Computational Creativity*, pp. 54–59, Mexico City, Mexico, (2011).

[23] Scott R Turner, *The creative process: a computer model of storytelling and creativity*, Erlbaum, Hillside, NJ, 1994.