Stepping Back to Progress Forwards: Setting Standards for Meta-Evaluation of Computational Creativity

Anna Jordanous
Centre for e-Research, Department of Digital Humanities
King's College London
26-29 Drury Lane, London WC2B 5RL, UK
anna.jordanous@kcl.ac.uk

Abstract

There has been increasing attention paid to the question of how to evaluate the creativity of computational creativity systems. A number of different evaluation methods, strategies and approaches have been proposed recently, causing a shift in focus: which methodology should be used to evaluate creative systems? What are the pros and cons of using each method? In short: how can we evaluate the different creativity evaluation methodologies? To answer this question, five meta-evaluation criteria have been devised from cross-disciplinary research into good evaluative practice. These five criteria are: *correctness; usefulness; faithfulness as a model of creativity; usability of the methodology; generality*. In this paper, the criteria are used to compare and contrast the performance of five various evaluation methods. Together, these meta-evaluation criteria help us explore the advantages and disadvantages of each creativity evaluation methodology, helping us develop the tools we have available to us as computational creativity researchers.

Introduction

Computational creativity evaluation repeatedly appears as a theme in the calls for papers for the ICCC conference series. Such emphasis underlines the growing importance of evaluation to the computational creativity research community.

For transparent and repeatable evaluative practice, it is necessary to state clearly what standards/methods are used for evaluation (Jordanous 2012a). Despite, or perhaps because of, a lack of creativity evaluation being employed in the computational creativity research community until recently (Jordanous 2011), a number of creativity evaluation strategies have been proposed in recent years (Pease, Winterstein, and Colton 2001; Ritchie 2007; Colton et al. 2010; Colton, Charnley, and Pease 2011; Jordanous 2012b). Herein lies a decision for a computational creativity researcher: which evaluation strategy should be adopted to evaluate computational creativity systems? What are the benefits and disadvantages of each?

Such questions have not previously been examined to any detailed extent in computational creativity research. In various other research fields, though, issues around 'evaluating evaluation', or *meta-evaluation*. have been considered in some detail. Meta-evaluation has been considered from philosophical and more practical standpoints. As a burgeoning research community, computational creativity researchers can learn from such considerations, as they apply to our own research efforts.

This paper proposes five standards for meta-evaluation of creativity evaluation methodologies, informed by the wider literature and by evaluative practices outside of the computational creativity field. These standards are offered as factors for assessment and comparison of creativity evaluation methodologies, to help us develop good evaluative practice in computational creativity research.

The five meta-evaluation standards are applied to a case study on creative system evaluation, comparing different evaluation methodologies against each other. Results are reported below. It is proposed that these five standards should help guide us in refining our work on computational creativity evaluation, as we progress in the development of this important area of computational creativity research.

The need to evaluate creativity evaluation

We have an intuitive but tacit understanding of the concept of creativity that we can access introspectively (Kaufman 2009; Jordanous 2012a). For comparative purposes and methodical, transparent evaluation, this intangible understanding is not sufficient to help us identify and learn from our successes and failures in computational creativity research.

To solve the problem of how to evaluate creative systems, various evaluation methodologies or strategies have been offered including the tests offered by Pease, Winterstein, and Colton, Ritchie's empirical criteria, the creative tripod model, the FACE model and the SPECS methodology (Pease, Winterstein, and Colton 2001; Ritchie 2007; Colton et al. 2010; Colton, Charnley, and Pease 2011; Jordanous 2012b, respectively).[1] But which should computational creativity researchers use?

One should note here that we are unlikely to find one single fully-specified, detailed, step-by-step methodology to suit all types of creative system. What we can do is to understand the strengths and weaknesses of different methodologies. Through trial, application of and comparison between different methodologies, refine and develop our eval-

[1] See Jordanous 2012a for full discussion of these methodologies and strategies.

uation strategies within computational creativity so that we can mutually learn from our advances and mistakes; the very essence of what evaluation offers researchers, after all.

How can these methodologies be compared against each other? Reviewing various features of the methodologies and comparing them against each other helps us to learn through comparison. Below, five meta-evaluation standards are identified for comparison and evaluation of creativity evaluation methodologies. These five meta-evaluation standards are drawn from cross-disciplinary reviews of evaluative practice. The meta-evaluation standards are applied in a practical case study, reported below. From this application of the standards, we can appreciate the strengths and weaknesses of each creativity evaluation methodology, guiding us in our evaluative choices when developing computational creativity research. With these meta-evaluation criteria, we can now compare evaluative results obtained through different methods and discuss how useful each of these evaluations are to the computational creativity researcher. Gathering effective evaluative feedback, using solidly developed evaluation methodologies, assists further computational creativity research development and helps identify more clearly the contributions to knowledge made by our research.

Criteria for meta-evaluation of creativity evaluation methodologies

Criteria for evaluation should be clearly stated and justified (Jordanous 2012a). This theme also applies to meta-evaluation criteria for comparing various creativity evaluation methodologies.

Certain areas suggest themselves as meta-evaluation criteria for assessing creativity evaluation methodologies, such as the accuracy and usefulness of the feedback to a researcher, or ease of applicability.

Pease, Winterstein, and Colton (2001) identify two candidate meta-evaluation criteria:

> 'Firstly, to what extent do they reflect human evaluations of creativity, and secondly, how applicable are they?' (Pease, Winterstein, and Colton 2001, p. 9)

More recently, Pease has suggested the set of {generality, usability, faithfulness, value of formative feedback} as candidate criteria (Pease, 2012, personal communications). In relevant literature on evaluation and related literature on proof of hypotheses in scientific method, other contributions could also be used as criteria for measuring the success of computational creativity evaluation methodologies, as outlined below.

Criteria for testing scientific hypotheses and explanatory theories Sloman (1978) outlined seven types of 'interpretative aims of *science*' (Sloman 1978, p. 26, my emphasis added), of which the third aim is the forming of explanatory theories for things we know exist. In the context of this current work, an example of the explanatory theories mentioned in the third aim would be a theory that allows us to explain if or why a computational creativity system is creative. Ten criteria were offered by Sloman (1978) as criteria for comparison of explanatory theories.

> 'a good explanation of a range of possibilities should be definite, general (but not too general), able to explain fine structure, non-circular, rigorous, plausible, economical, rich in heuristic power, and extendable.' (Sloman 1978, p. 53)

Within these criteria there is some significant interdependence and Sloman advises that the criteria are best treated as a set of inter-related criteria rather than distinct yardsticks, with some criteria (such as plausibility, generality and economy) to be used with caution. This may help to explain why Sloman's list of criteria is longer than others mentioned in this Section.

Thagard (1988) defined a 'good theory' as 'true, acceptable, confirmed' (Thagard 1988, p. 48). These criteria were later expressed in the form of 'the criteria of consilience, simplicity of analogy' (Thagard 1988, p. 99) as essential criteria for theory evaluation:

- *Consilience* - how comprehensive the theory is, in terms of how much it explains.
- *Simplicity* - keeping the theory simple so that it does not try to over-explain a phenomenon. Thagard mentions in particular that a theory should not try to 'achieve consilience by means of ad hoc auxiliary hypotheses' (Thagard 1988, p. 99). In other words, the main explanatory power of the theory should map closely to the main part of that theory, without needing extensive correction and supplementation.
- *Analogy* - boosting the 'explanatory value'(Thagard 1988, p. 99) of a theory by enabling it to be applied to other demands. This is especially appropriate where theories can be cross-applied in more established domains where knowledge of facts is more developed.

Guidelines for good practice in research evaluation Suggestions for good practice in performing evaluation in research can be interpreted as criteria that identify such good practice. For example, in his 'Short Course on Evaluation Basics', John W. Evans identifies four 'characteristics of a good evaluation':[2] a good evaluation should be objective, replicable, generalisable and as 'methodologically strong as circumstances will permit'. In considering what constitutes good evaluation practice, the MEERA website ('My Environmental Education Evaluation Resource Assistant')[3] describes 'good evaluation' as being: 'tailored to your program ... crafted to address the specific goals and objectives [of your program'; '[building] on existing evaluation knowledge and resources'; inclusive of as many diverse viewpoints and scenarios as reasonable; replicable; as unbiased and honest as possible; and 'as rigorous as circumstances allow'. From a slightly different perspective on research evaluation, the European Union FP6 Framework Programme describes how FP6-funded projects are evaluated in terms of three criteria:

[2] http://edl.nova.edu/secure/evasupport/evaluationbasics.html, last accessed Feb 2014.

[3] All quotes from the MEERA website are taken from http://meera.snre.umich.edu/plan-an-evaluation/evaluation-what-it-and-why-do-it#good, last accessed Feb 2014.

a project's *rationale* relative to funding guidelines and resources; *implementation* effectiveness, appropriateness and cost-effectiveness; and *achievements* and impact of contributions of objectives and outputs.

Dealing with subjective and/or fuzzy data: Blanke's specificity and exhaustivity In computational creativity evaluation, the frequency of data being returned is low and the correctness of that data is generally subjective and/or fuzzy in definition, rather than being discretely categorisable as either correct or incorrect, or as either present or missing. Blanke (2011) looked at how to evaluate the success of a methodology for measuring aspects like precision and recall, in cases where the results being returned were somewhat difficult to pin down to exact matches due to fuzziness in what could be returned as a correct result. The specific case Blanke considered was in XML retrieval evaluation, where issues such as hierarchical organisation and overlap of elements, and the identification of what was an appropriate part of an XML document to return, caused problems with using precision and recall measures. There was also an issue with relatively low frequencies in what was being returned.

As an evaluation solution, Blanke (2011) proposed *component specificity* and *topical exhaustivity*, following from Kazai and Lalmas (2005). Exhaustivity 'is measured by the size of overlap of query and document component information' (Blanke 2011, p. 178). Specificity 'is determined by counting the rest of the information in the component [of an XML document] that is not about the query' (Blanke 2011, p. 178), such that minimising such information will maximise the specificity value, as more relevant content is returned.

Identifying meta-evaluation criteria

Drawing all the above contributions together, five criteria can be identified for meta-evaluation of computational creativity evaluation methodologies. These are presented here, with relevant points from the comments above being grouped under the most relevant criterion, as far as possible. Some overlap across criteria is acknowledged, for example Thagard's *analogy* criterion can be interpreted as being concerned with both 'usefulness' and 'generality'.

- **Correctness: how accurately and comprehensively the evaluation findings reflect the system's creativity.**
  - MEERA's *honesty of evaluation* criterion.
  - MEERA's *inclusiveness of diverse relevant scenarios* criterion.
  - Evans' *objectiveness* criterion.
  - MEERA's *avoidance of bias in results* criterion.
  - Sloman's *definiteness* criterion.
  - Sloman's *rigorousness* criterion.
  - Sloman's *plausibility* criterion.
  - Thagard's *consilience* criterion.
  - Blanke's *exhaustivity* criterion.
  - Evans' *methodological strength* criterion.
- **Usefulness: how informative the evaluative findings are for understanding and potentially improving the creativity of the system.**
  - Pease's *value of formative feedback* criterion.
  - FP6's *rationale*, *implementation* and *achievements* criteria.
  - Sloman's *heuristic power* criterion.
  - Thagard's *analogy* criterion.
- **Faithfulness as a model of creativity: how faithfully the evaluation methodology captures the *creativity* of a system (as opposed to other aspects of the system).**
  - Pease, Winterstein, and Colton (2001)'s *reflection of human evaluations of creativity* criterion.
  - Pease's *faithfulness* criterion.
  - MEERA's *tailoring of the method to specific goals and objectives* criterion.
  - Blanke's *specificity* criterion.
- **Usability of the methodology: the ease with which the evaluation methodology can be applied in practice, for evaluating the creativity of systems.**
  - Pease, Winterstein, and Colton (2001)'s *applicability* criterion.
  - Pease's *usability* criterion.
  - Evans' *replicability* criterion.
  - MEERA's *replicability* and *rigorousness of a methodology* criteria.
  - Sloman's *non-circularity* criterion.
  - Sloman's *rigorous* and *explicitness* criteria (in how to apply the methodology).
  - Sloman's *economy of theory* criterion.
  - Thagard's *simplicity* criterion.
- **Generality: how generally applicable this methodology is across various types of creative systems.**
  - Pease's *generality* criterion.
  - MEERA's *inclusiveness of diverse relevant scenarios* criterion.
  - Evans' *generalisability* criterion.
  - Sloman's *generality* criterion.
  - Sloman's *extendability* criterion.
  - Thagard's *analogy* criterion.

Applying the criteria: a case study

Now we have identified these five meta-evaluation criteria, we can use them to evaluate the performance of computational creativity evaluation methodologies.

Previously, three different musical improvisation computer systems were evaluated using various computational creativity evaluation methodologies, to compare how creative each system was (Jordanous 2012a; 2012b). The task in this current work is to consider how well the creativity evaluation methodologies performed for this assessment.

For an independent assessment of the relative performance of the evaluation methodologies, external evaluation was sought to consider and perform meta-evaluation

on five key existing evaluative approaches (Ritchie 2007; Colton 2008; Colton, Charnley, and Pease 2011; Jordanous 2012b, surveys of human opinion). The invited external evaluators were the key researchers involved in creating the musical improvisation systems examined in the above-mentioned creativity evaluation case study(Jordanous 2012a): Al Biles (GenJam) and George Lewis (Voyager). Bob Keller was also invited because of his research into and development of a related musical improvisation system, the Impro-Visor system (Gillick, Tang, and Keller 2010).[4] Evaluators were asked to view all the evaluative feedback obtained. They were then asked to give their opinions (as developers of musical improvisation systems) on various aspects of each methodology and on the results obtained.

Below, the methodology used for the meta-evaluation is briefly described, and the obtained meta-evaluations are reported and discussed. Fuller details can be found in Jordanous (Jordanous 2012a).

Methodology for obtaining external evaluation

Each external evaluator was given a feedback sheet reporting the evaluation feedback obtained for their system from each creativity evaluation methodology being investigated: Ritchie's criteria; Colton's creative tripod; survey of human opinion; the FACE model; and SPECS+cc. (N.B. *SPECS+cc* is used here to indicate the use of Jordanous's SPECS methodology with the 14 creativity components (Jordanous 2012a) as the adopted definition of creativity, as recommended (Jordanous 2012b).)

For each methodology, the sheets also included brief comparisons between systems according to the systems' evaluated creativity. An example of these feedback sheets, given in (Jordanous 2012a, Appendices), presents the sheet provided to Al Biles to report the evaluation results for GenJam. A similar set of feedback was prepared and sent to George Lewis as evaluative feedback relating to Voyager. Methodologies were presented under anonymous identifiers in the feedback sheet to avoid any bias from being introduced, as far as possible.

Evaluators were first asked if they had any initial comments on the results. They were then asked to provide full feedback for each methodology in turn, on the five criteria derived above. They looked at all five criteria for the current methodology and then were asked for any final comments on that methodology before moving onto the next methodology. Methodologies were presented to the evaluators in a randomised order, to avoid introducing any ordering bias.

For each criterion, questions and illustrating examples were composed to present the criterion in a context appropriate for computational creativity evaluation. These questions and examples, listed below, were put to external evaluators to gather their feedback on each criterion as meta-evaluation of the various evaluation methodologies.

- **Correctness:**
  - How correct do you think these results are, as a reflection of your system?
  - For example: are the results as accurate, comprehensive, honest, fair, plausible, true, rigorous, exhaustive, replicable and/or as objective as possible?
- **Usefulness:**
  - How useful do you find these evaluation results, as an / the author of the system?
  - For example: do the results provide useful information about your system, give you formative feedback for further development, identify contributions to knowledge made by your system, or give other information which you find helpful?
- **Faithfulness as a model of creativity:**
  - How faithfully do you think this methodology models and evaluates the creativity of your system?
  - For example: do you think the methodology uses a suitable model(s) of creativity for evaluation, does the methodology match how you expect creativity to be evaluated, how specifically does the methodology look at creativity (rather than other evaluative aims)?
- **Usability of the methodology:**
  - How usable and user-friendly do you think this methodology is for evaluating the creativity of computational systems?
  - For example: would you find the methodology straightforward to use if wishing to evaluate the creativity of a computational creativity system (or systems), is the methodology stated explicitly enough to follow, is the method simple, could you replicate the experiments done with this methodology in this evaluation case study?
- **Generality:**
  - How generally do you think this methodology can be applied, for evaluation of the creativity of computational systems?
  - For example: can the methodology accommodate a variety of different systems, be generalisable and extendable enough to be applied to diverse examples of systems, and/or different types of creativity?

For each criterion, evaluators were asked to rate the system's performance on a 5 point Likert scale (all of a format ranging from positive extreme to negative extreme, such as: [Extremely useful, Quite useful, Neutral, Not very useful, Not at all useful]). They could also add any comments they had for each criterion.

Evaluators were asked about the correctness and usefulness of the methodology's results, before learning how the methodology worked. This gave the advantage of being able to hear the evaluators' opinions considering the feedback results in isolation, without any influence from how the results were obtained. Nonetheless, the process by which a product was generated is important to consider alongside that product, for a more rounded and informed evaluation (Rhodes

[4] The author of one evaluated systems (GAmprovising) was not included, due to being the author of one of the evaluation methods being examined (and the researcher conducting this work).

1961). Evaluators were given details on how that methodology worked after evaluating the correctness and usefulness criteria. They were then asked to provide feedback for the final three criteria (faithfulness, usability and generality). The details provided to explain each methodology are reproduced in Jordanous (Jordanous 2012a, Appendices).[5]

Finally, evaluators were asked to rank the evaluation methodologies according to how well they thought the methodologies evaluated the creativity of their system overall. Although the formative feedback is, again, probably more useful in terms of developing the various methodologies, it was interesting to see evaluators' opinions on how the methodologies compared to each other. The rankings, completed by Al Biles and Bob Keller, are reported in Table 1. At this point, evaluators were also given a change to add any final comments, before finishing the study. Al Biles completed a full evaluation of all methodologies and (due to time constraints) George Lewis provided evaluations of two methodologies: Colton's creative tripod and the SPECS+cc methodology. Bob Keller also provided comments on some aspects of all methodologies.

Results and discussion of meta-evaluation

Al Biles summarised the meta-evaluation of the five different methodologies with: *'Five very different approaches, and each bring something to the table.'* In the comparisons between methodologies and the overall rankings listed in Table 1, SPECS+cc was either considered the best methodology overall (ahead of the creative tripod) or the second best (behind Ritchie's criteria) for evaluating a system's creativity. The more useful information, though comes from the more detailed formative feedback and comments rather than a single summative ranking as given in Table 1.

SPECS+cc was evaluated by both Biles and Lewis, with some additional comments from Keller. SPECS+cc generated 'extremely useful' and 'quite correct results', in both of the main evaluators' opinions. One evaluator found SPECS+cc to be an 'extremely faithful' model of creativity, though the other was 'neutral' on this matter. While one evaluator found SPECS+cc 'quite user-friendly', the other questioned how user-friendly the SPECS+cc methodology would be, given the steep learning curve in understanding the components. In terms of generality, evaluators disagreed on how generally SPECS+cc could be applied, further comments illustrated how methods like SPECS+cc were more appropriate for taking into account other system goals, compared to more limited views on creativity such as in the FACE model. Biles and Keller in particular commented on the lack of accommodation of other system goals in the FACE model, though it is to be acknowledged that such accommodation does not form one of the goals of the FACE model and is more of an unintended but useful consequential result in models such as SPECS+cc.

FACE was placed third in the overall rankings by Biles and last by Keller. Biles, the main evaluator for FACE, found the results generated by FACE to be 'completely correct', but gave a neutral opinion (neither positive nor negative) on the usefulness of FACE model feedback, the generality of the FACE model across domains and the faithfulness of the FACE model as a model of creativity. FACE was deemed 'quite user-friendly' due to its simplicity; this opinion was repeated, more strongly, for the other creativity evaluation framework Colton was involved in, the creative tripod. Lewis and Biles both evaluated the tripod; they disagreed as to whether the tripod would be generally applicable across many domains, and also as to how faithfully the tripod modelled creativity. Both evaluators agreed, however, that the feedback from the tripod was 'extremely useful' and either 'completely correct' or 'quite correct'. Biles ranked the creative tripod as the second best creativity evaluation methodology overall, though Keller placed it last.

Ritchie's criteria methodology was fully evaluated by Biles. Biles found the criteria to produce 'quite correct', 'quite useful' feedback that was 'quite faithful to creativity' (despite raising issues with enforced simplifications of the data due to the boolean rather than continuous nature of the feedback). Biles was 'neutral' on the usability of applying the criteria for creativity evaluation and on their generality, questioning how the generic terminology used to solicit ratings of typicality and value could be applied to different domains successfully. Keller considered Ritchie's criteria to be the best methodology overall for creativity evaluation, though Biles gave it a middling ranking.

The opinion survey was ranked overall to be the fourth best methodology out of the five. It received a few negative comments from Biles, the main evaluator for this system, despite Biles noting that 'nothing is simpler than just ... asking whether something is creative or not' and that the survey solicited spontaneous,'unadulterated' opinions rather than restructuring the feedback (though Biles also noted that the tripod feedback was clearer than the survey feedback due to its more structured presentation). Biles was guided in a number of comments by an observation that the opinion survey sacrificed reliability/consistency of results for greater validity in terms of the personal qualitative feedback. He thought that the survey approach could be applied 'quite generally' and was 'quite user-friendly' and 'quite faithful' to what it means to be creative. The success of this methodology would depend on the type of person participating, and whether they were clear on what 'creative' means. Given that the GenJam system has been publicly presented many times before, though, Biles felt he learned nothing new from the feedback from the survey, unlike the other methodologies. He was 'neutral' on the correctness of the methodology, confirming observations made in Jordanous 2012a that human opinion cannot be relied on as a 'ground truth' to measure evaluations against, due to varying viewpoints.

[5] It is worth noting that methodologies may well perform differently against the five criteria when applied to different systems (a meta-application of the generality criterion?) The evaluators cannot be expected to give rigorous feedback on the potential of the methodologies in evaluating *any* possible type of system, and we should refrain from drawing too-broad conclusions from their feedback. Nonetheless, with careful consideration of the evaluators' feedback, we gain valuable insights on the methodologies.

Table 1: Judges were asked to rank the methodologies according to how well overall they thought the methodologies evaluated the systems' creativity:

Position	Al Biles	Bob Keller
1st (best)	SPECS+cc	Ritchie's criteria
2nd	Creative Tripod	SPECS+cc
3rd	Ritchie's criteria	FACE
4th	Opinion survey	Opinion survey
5th (worst)	FACE	Creative Tripod

Comparing and contrasting methodologies

Five meta-evaluation criteria have now been identified for meta-evaluation of creativity evaluation methodologies and have been used for evaluation by external evaluators, as reported above. Next, the criteria were applied for further analysis of all the methodologies investigated earlier in this paper, using the full findings from the Jordanous (2012a) case study evaluating the creativity of musical improvisation systems. Such considerations on the methodologies allow us to compare if, and how, a particular evaluation methodology marks a development of our evaluation 'toolkit' as computational creativity researchers. Here, the considerations are focused towards evaluating how well the SPECS+cc methodology (Jordanous 2012a) performed, to gain feedback as to how to improve SPECS+cc and what its strengths were in comparison to other methods. The considerations below also complement the evaluative case study findings by accounting for more detailed information and observations that may not have been detected by the external evaluators, but which should still be considered.

Correctness Showing that human opinion cannot necessarily be relied on as a ground truth, even on a large scale, some participants in opinion surveys admitted that they were likely to be evaluating the systems based on how highly they rated a system's performance overall rather than specifically how creative they thought it was, which would affect the overall correctness of the results of evaluations from the human opinion survey.

SPECS+cc performed better than Ritchie's criteria for correctness. Although Ritchie's 18 criteria have a comprehensive coverage of observations over the products of the system, criteria evaluation is based solely on the products of the creative system, not accounting for the system's process, or observations on the system or how it interacted with its environment. Colton's tripod model was found to be reasonably accurate in terms of identifying and evaluating important aspects in the case study, but it has disregarded aspects such as social interaction, communication and intention, which have been shown to be very important in understanding how musical improvisation creativity is manifested (Jordanous and Keller 2014).

It should be noted that 'correctness' does not imply that the results from evaluation match common human consensus as a 'ground truth', or 'right answer'; Jordanous (Jordanous 2012a) demonstrated that these are not reliable goals in creativity evaluation. Instead, correctness is concerned with how appropriate the feedback is and how accurately and realistically the feedback describes the system.

Usefulness The methodologies differed in the amount of feedback generated through evaluation. A fairly large volume of qualitative and quantitative feedback was returned through the application of SPECS+cc. This is unlike Ritchie's criteria which only returned a set of 18 Boolean values, one for each criterion, with some interpretation effort needed to understand how each criterion influences creativity within the system.[6] Colton's creative tripod generated feedback for 3 components, rather than 14 components, so was shorter than SPECS+cc. The human opinion surveys generated similar quantities of feedback to SPECS+cc, from more people but a shallower level of detail.

The human opinions surveys returned less detailed feedback than SPECS+cc, which generated a large amount of detailed formative feedback. The opinion surveys' feedback also often concentrated on aspects of the systems other than its creativity, according to participant feedback (Jordanous 2012a).

Ritchie's criteria returned a set of 18 boolean values rather than any formative feedback, in a fairly opaque form given the formal abstraction of the criteria specification; if there were no output examples, Ritchie's criteria would not generate any feedback at all, even based on other observations about the system. Colton's creative tripod returned information at the same level of detail as SPECS+cc per component/tripod quality, but less information overall, as several useful components of SPECS+cc were overlooked because they did not map onto the set of tripod aspects.

Faithfulness as a model of creativity Participant feedback for the human opinion surveys acknowledged that evaluations may have related more to the quality of the system, not its creativity, with several participants requesting a definition of creativity to refer to when evaluating how creative the systems were (Jordanous 2012b). The SPECS methodology requires researchers to base their evaluations on a researched and informed understanding of creativity that takes into account both domain-specific and domain-independent aspects of creativity. In this way it is the only methodology that directly accounts for specific informed requirements for creativity in a particular domain. Human opinion surveys would acknowledge this but only tacitly, without these requirements necessarily being identifiable or explainable. Although the parameters and weights in Ritchie's criteria could be customised to reflect differing requirements for creative domains, in practice no researchers have attempted this

[6] One reviewer of this paper pointed out that Ritchie (2007) also briefly considered how his criteria could be adapted to return measurements of each criterion in the range [0,1], rather than Boolean values, although Ritchie's main presentation of the criteria is as statements which generate Boolean values. This alternative usage gives slightly more information, but the issues of interpreting these criteria's contribution to overall creativity still remain.

when applying Ritchie's criteria, probably due to the formal and abstracted presentation of the criteria. In Colton's creative tripod, all three tripod qualities are treated equally in previous examples (including those in Colton (2008)) regardless of their contribution in a specific creative domain and no further qualities can be introduced into the tripod framework.

Usability of the methodology Less information needed to be collected for Colton's creative tripod than for the other methodologies, taking less time to collect. Coupled with the informal nature of performing creativity evaluation with the tripod framework, Colton's creative tripod emerged as the most easy-to-use of the methodologies evaluated. Data collection for the other methodologies was of a similar magnitude, although data analysis for Ritchie's criteria was slightly more involved and more specialist than the other methodologies, requiring a specific understanding of the criteria.

Feedback reflected on the volume of data generated by using the components as a base model of creativity, as recommended for SPECS. If SPECS is applied without using the Jordanous (2012b) components as the basis for the adopted definition of creativity, then SPECS becomes more involved and more demanding in terms of researcher effort, negatively affecting its usability. Hence the recommendation in Jordanous (2012b) for using the components within SPECS (i.e. SPECS+cc) becomes further strengthened.

One issue is with who/what performs evaluation, and what effect that has on how usable the evaluation methodology. Using external evaluators increases the time demands of the experiment in the human opinion surveys, as this requires studies to be carried out and introduces extra work to be done such as planning experiments for participants or applying for ethical clearance for conducting experiments with people. While the use of external evaluators is not a formal requirement for the SPECS+cc methodology - indeed evaluation can be performed using quantitative tests rather than subjective judgements if deemed most appropriate - the accompanying commentary to SPECS+cc strongly encourages researchers to use independent evaluation methods in order to capture more independent and unbiased results (Jordanous 2012b). In the application of SPECS+cc that is being reviewed here, external judges were consulted to give feedback on the creative systems being evaluated. Hence SPECS+cc in this case is subject to similar criticisms, in terms of ease of use, as when conducting opinion surveys. These extra demands are not necessarily encountered when performing evaluation as recommended using Colton's tripod, Ritchie's criteria, or FACE evaluation, where no specific demands or recommendations are made for evaluation to be performed independently of the project team behind the creative software. It is important to acknowledge, though, that should independent evaluation be sacrificed in order to make an evaluation methodology more useful, there is a worrying knock-on effect, in terms of potential biases being introduced if evaluation is not being performed by independent evaluators.

Generality SPECS+cc, Colton's tripod and to some extent, Ritchie's criteria and the human opinion surveys, could all be applied to different types of system, providing that the system produces the appropriate information relevant to the individual methodologies.[7] Ritchie's criteria cannot be applied to systems that produce no tangible outputs, making this approach less generally applicable across creative systems. There is also some question of whether opinion surveys could be carried out for evaluating all types of creativity, particularly where creativity is not manifested outwardly in production of output, affecting the generality of opinion surveys.

Overall comparisons Considering all the observations made in this paper from the perspective of the five meta-evaluation criteria presented in this paper, SPECS+cc performed well in comparison with the other evaluation methodologies on its faithfulness in modelling creativity. SPECS+cc also performed better than Ritchie's criteria for usefulness and correctness and produced larger quantities of useful feedback than Colton's creative tripod (because less information was collected for Colton's creative tripod). A consequence of the information collection meant that Colton's creative tripod was the easiest to use of the methodologies evaluated.

Somewhat counterintuitively, all the methodologies were more likely to generate correct results compared to the surveys of human opinion. A number of participants in the opinion surveys reported that they evaluated systems based on factors other than creativity, due to difficulties in evaluating creativity of the Case Study systems without a definition of creativity to refer to. There is also some question of whether human opinion surveys could be carried out for evaluating all types of creativity (particularly where creativity is not manifested outwardly in copious production of output); this affects the general applicability of using opinion surveys. Reliance on the existence of output examples also affects the usability and generalisability of Ritchie's criteria.

Conclusions

Several evaluation methods were applied to three musical improvisation systems. Human opinion was consulted to try and capture a 'ground truth' for creativity evaluation (Zhu, Xu, and Khot 2009). Four key existing methodologies for computational creativity were also applied (Ritchie 2007; Colton 2008; Colton, Charnley, and Pease 2011; Jordanous 2012b, Ritchie's criteria, the Creative Tripod, the FACE model and the SPECS+cc methodology, respectively). Results were compared; it was noted that few 'right answers' or 'ground truths' for creativity were found.

For the purposes of progressing in research, learning from advances and improving what has been done, how well did each evaluation methodology perform? To assist in answering this question, external evaluation was solicited from the authors of the evaluated musical improvisation systems and one other researcher with interests in creative musical improvisation systems.

[7] This is illustrated further in Case Study 2 in Jordanous 2012a.

Five criteria were identified from relevant literature sources for meta-evaluation of important aspects of the evaluation methodologies:

- Correctness
- Usefulness
- Faithfulness as a model of creativity
- Usability of the methodology
- Generality

The methodologies were compared based on the external evaluators' feedback concerning the evaluations performed on their system and the comparative feedback generated by each methodology considered so far. Further comments could be made using the meta-evaluation criteria, based on detailed study of the methodologies themselves.

These results are too small in number to be a comprehensive evaluation but they do help to give us some feedback on the compared methodologies. The results showed that SPECS+cc and Ritchie's empirical criteria compared favourably to the other methodologies overall. SPECS+cc performed well on most of the five meta-evaluation criteria, though the volume of data produced by SPECS+cc raised questions on SPECS+cc's usability compared to more succinct presentations. Colton's creative tripod was the easiest to use although there were some concerns about the generality of the tripod across creative domains and its faithfulness as a general model of creativity. Ritchie's criteria were considered accurate but there were usability issues with the abstract nature of the criteria and accompanying function definitions. The FACE model was considered quite user friendly but perhaps limited in how it could incorporate aspects of creativity that were important to the system domain but outside of the face model. Each of the evaluation methodologies proved to be an improvement (in at least some ways) over the approach of simply asking people's opinions on how creative the systems were.

The development of creativity evaluation methods is clearly a key current area of interest in the computational creativity research community, as partly illustrated by the prominent inclusion of requests for papers on evaluation, in the call for papers for ICCC 2014. The five meta-evaluation criteria offered in this paper are taken from a cross-disciplinary review of good practice in evaluation of areas relevant to computational creativity research. These five criteria help us to contrast different evaluation methodologies against each other

Acknowledgments

Thanks to Alison Pease, Steve Torrance and Nick Collins for helpful comments during the formulation of these ideas. Also thanks to Al Biles, Bob Keller and George E. Lewis for willingly offering their time and helpful comments as evaluators for this work, and to the three anonymous reviewers for their useful remarks on the original version of this paper.

References

Blanke, T. 2011. *Using Situation Theory to evaluate XML retrieval*. Dissertations in Database and Information Systems. Heidelberg, Germany: IOS Press.

Colton, S.; Gow, J.; Torres, P.; and Cairns, P. 2010. Experiments in objet trouvé browsing. In *Proceedings of the International Conference on Computational Creativity*.

Colton, S.; Charnley, J.; and Pease, A. 2011. Computational Creativity Theory: The FACE and IDEA descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity*, 90–95.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Proceedings of AAAI Symposium on Creative Systems*, 14–20.

Gillick, J.; Tang, K.; and Keller, R. M. 2010. Machine learning of jazz grammars. *Computer Music Journal* 34(3):56–66.

Jordanous, A., and Keller, B. 2014. What makes musical improvisation creative? *Journal of Interdisciplinary Music Studies* Forthcoming.

Jordanous, A. 2011. Evaluating evaluation: Assessing progress in computational creativity research. In *Proceedings of the Second International Conference on Computational Creativity (ICCC-11)*.

Jordanous, A. 2012a. *Evaluating Computational Creativity: A Standardised Procedure for Evaluating Creative Systems and its Application*. Ph.D. Dissertation, University of Sussex, Brighton, UK.

Jordanous, A. 2012b. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

Kaufman, J. C. 2009. *Creativity 101*. The Psych 101 series. New York: Springer.

Kazai, G., and Lalmas, M. 2005. Notes on what to measure in INEX. In *INEX 2005 Workshop on Element Retrieval Methodology*.

Pease, A.; Winterstein, D.; and Colton, S. 2001. Evaluating machine creativity. In *Proceedings of Workshop Program of ICCBR-Creative Systems: Approaches to Creativity in AI and Cognitive Science*, 129–137.

Rhodes, M. 1961. An analysis of creativity. *Phi Delta Kappan* 42(7):305–310.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:67–99.

Sloman, A. 1978. *The computer revolution in philosophy*. Hassocks, Sussex: Harvester Press.

Thagard, P. 1988. *Computational Philosophy of Science*. Cambridge, MA: MIT Press.

Zhu, X.; Xu, Z.; and Khot, T. 2009. How creative is your writing? a linguistic creativity measure from computer science and cognitive psychology perspectives. In *Proceedings of NAACL HLT Workshop on Computational Approaches to Linguistic Creativity (ACL)*, 87–93.