Evaluating Quality in Creative Systems

Graeme Ritchie
University of Aberdeen

Graeme Ritchie

Some Empiric [illegible] for Attributing Creativity t[illegible]ter Program.

Minds [illegible] *ines* 17 {1}, pp.67-99.

NOT TODAY

THIS TALK

- PART I : Introduction
  - The two meanings of “creative system”.
  - The importance of “quality”.
- PART II : Evaluation
  - Some ways to explore and assess “quality”.
- PART III : Other aspects
  - Speculations about wider issues.

PART Ⅰ : Introduction

What is meant by the term “creative system”?

“... software which exhibits behaviour that would be deemed creative if exhibited by a person”

Two common usages of “creative”

“loose” version

(a) an activity (e.g. painting, musical composition, writing poetry) which is regarded as inherently “creative”

(b) displaying particular skill or artistry or ingenuity

“strict” version

creative$_L$ vs. **creative**$_S$

- creative$_L$ but not creative$_S$:

  e.g. a music-composing program which generates simple and unoriginal melodies.

- creative$_S$ but not creative$_L$:

  e.g. a mathematical theorem generator which displays a high degree of originality and ingenuity.

- both creative$_L$ and creative$_S$:

  e.g. a poetry-writing program which generates subtle, profound and innovative verse

WHAT IS DISTINCTIVE ABOUT CREATIVE$_L$ SYSTEMS?

- The program generates “artefacts” (possibly abstract).
- The set of acceptable artefacts is not well-defined.
- The set of acceptable artefacts is very large (maybe infinite).
- Acceptability of an artefact depends on social, cultural, subjective, personal factors.
- There is a notion of artefacts “being good” (quality)

WHAT IS DIFFERENT ABOUT THIS? *(COMPARED TO SOFTWARE ENGINEERING)*

- No practical task
- No prior specification of “requirements”
- No precise definition of correctness
- No objective measure of success
- Different criteria for what counts as a “good” method of implementation

Why bother with evaluation?

- To see if we are making any progress.
- To see which methods/models give the best results.

TESTING “QUALITY”

“Quality” – how good are the artefacts produced by the system?

Why measure this?

- It is a first indication of success.
- Creativity (“strict”) is often seen as some combination of **quality** and **novelty**/innovation.

PART II - Evaluation

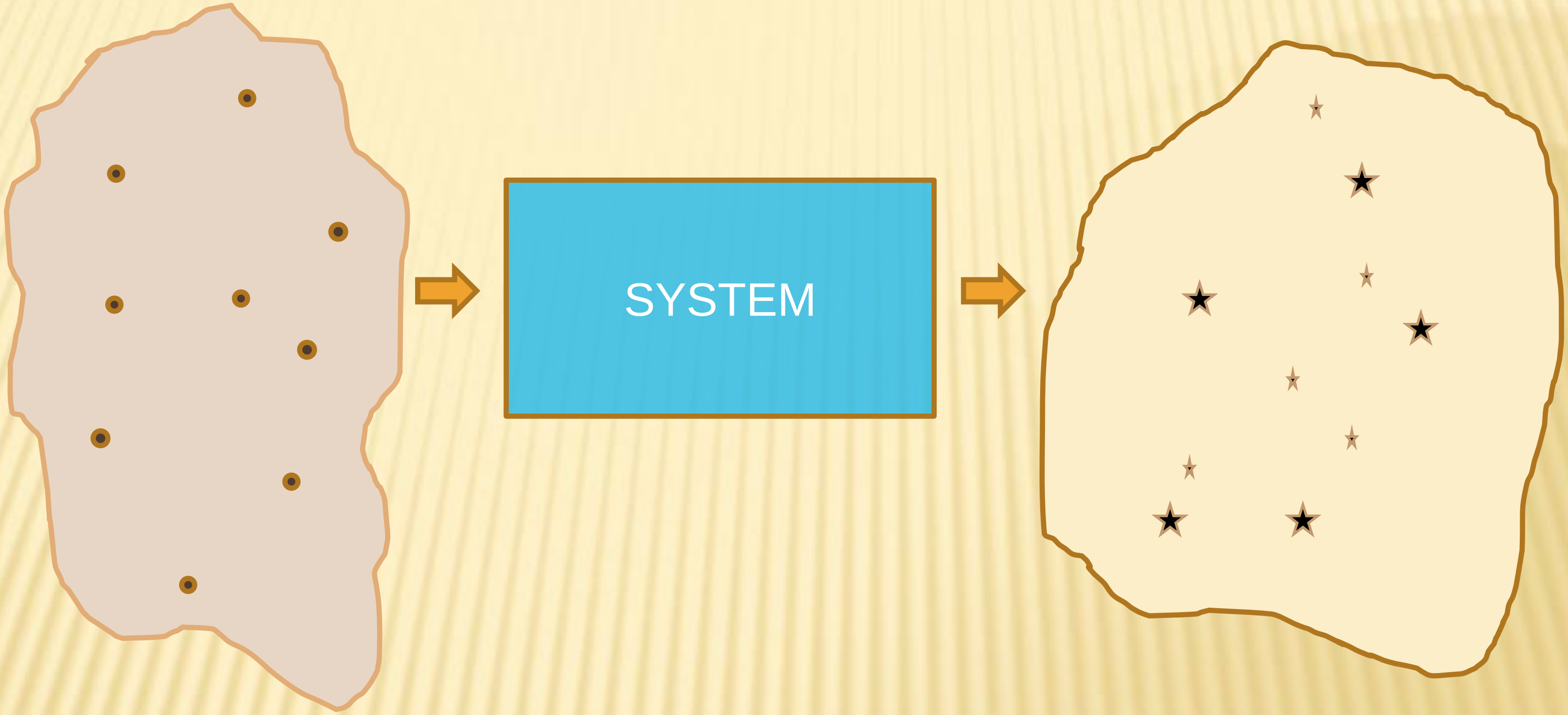
So how do we test programs like this?

DIFFERENT NOTIONS OF “TESTING”

- **Debugging:** informal, to find bugs, etc.
- **Formative evaluation:** organised and systematic tests to check progress during development.
- **Summative evaluation:** controlled assessment of the success of the final version of the system.

THE ASSUMED SITUATION

- We have an abstract model of generation (e.g. poems, melodies, pictures,...).
- There are parameters which can alter the behaviour of the model.
- We have implemented this model in a working program.
- The program has at least some of the parameters of our model.
- Formative evaluation indicates that the program is working well.



PARAMETERS

SYSTEM

ARTEFACTS

GENERATING THE OUTPUT

1. Re-creating known exemplars

Parameter values that reproduce existing artefacts?

This tests the accuracy of the model.

May be more suited to “formative” evaluation.

Not appropriate for all situations.

Success is not a sign of “strict” creativity.

GENERATING THE OUTPUT

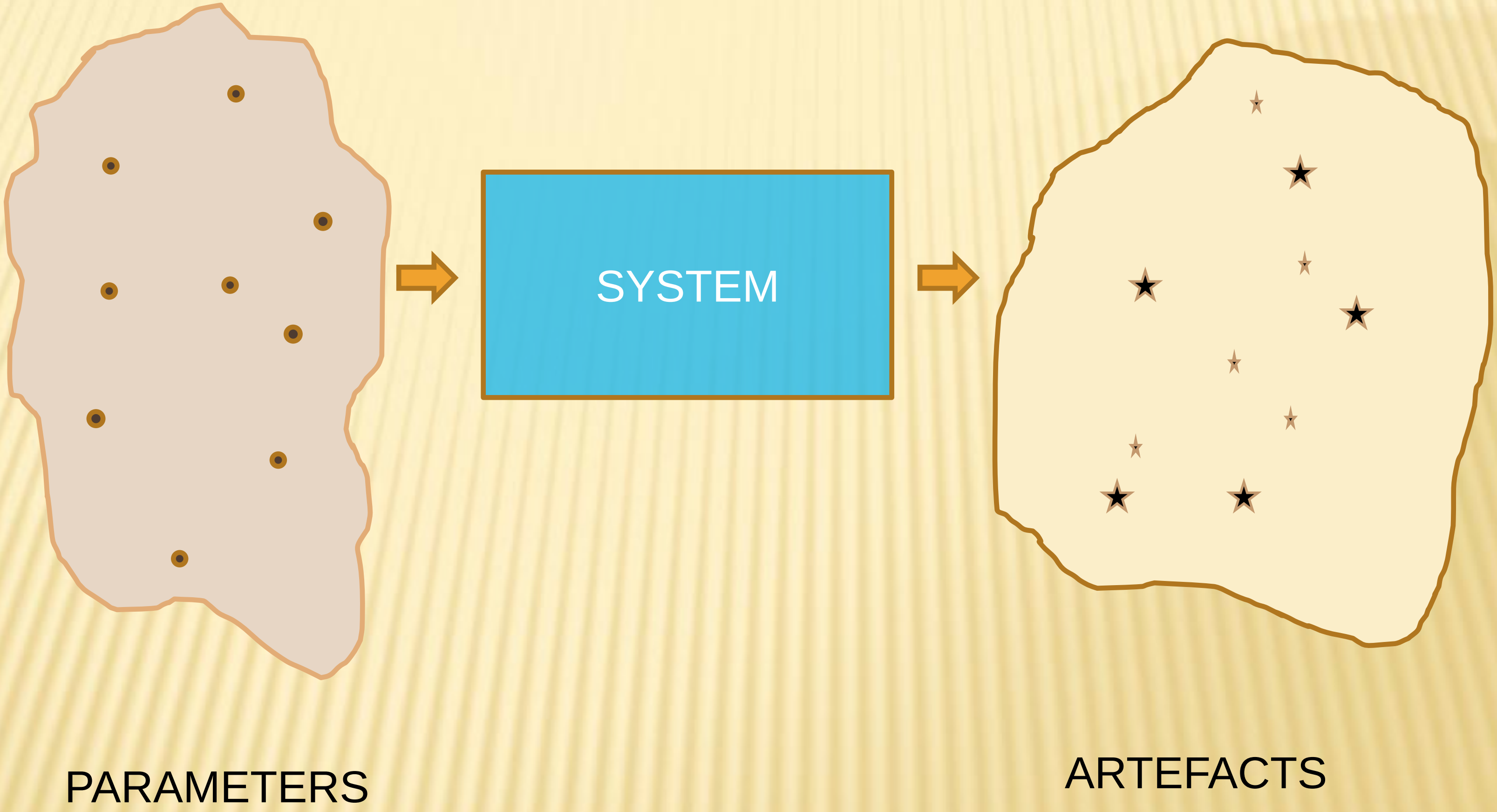
2. Exploring the neighbourhood of known exemplars

Parameter values that result in output *similar* to existing artefacts?

A possible “formative” evaluation method.

More realistic than aiming for exact replication of exemplars.

Success here may indicate a low level of “strict” creativity.



SYSTEM

PARAMETERS

ARTEFACTS

GENERATING THE OUTPUT

3. Exploring the parameter space

What happens if you make very small changes to the parameters?

Also a possible “formative” evaluation method.

More interesting than aiming for replication of exemplars.

Success here may indicate a low level of “strict” creativity.

GENERATING THE OUTPUT

4. Random sampling of parameter space

Choose parameter values randomly.

What is meant by “random” in this context?

Need good definition of parameter ranges.

More oriented towards summative evaluation.

Could show evidence of “strict” creativity.

GENERATING THE OUTPUT

5. Structured sampling of parameter space

Choose parameter values using a theoretical hypotheses.

You need to have some theoretical ideas.

More oriented towards summative evaluation.

Could be relevant to “strict” creativity.

SELECTING OUTPUT TO EVALUATE

? Assess all output from all evaluation runs of the system.

? Randomly select output items.

? Select output items in a structured way.

? Select output items in an ad hoc or subjective way.

HOW TO EVALUATE THE QUALITY OF THE OUTPUT

Since “quality” is subjective, two main approaches:

- Rating of output items by naive judges.
- Rating of output items by experts.

RATING BY NAIVE JUDGES

- A large number of judges (for statistical reasons).
- Judges **not** knowledgeable about your research.
- No technical/theoretical questions.
- Just a few questions.

PREPARING THE ITEMS (1)

As well as the program output, include:

- Control items: For comparison with system's output.
- Fillers : Irrelevant items.
- Practice: To demonstrate the task, or for practice.

PREPARING THE ITEMS (2)

What should the “control” items be?

- Human-created artefacts?
- Random instances of output medium?
- Items superficially similar to system’s output?
- System output with random parameters?
- Output from other systems?

CONSTRUCTING ITEM SETS

- Each item seen by a few judges.
- Each item seen by similar number of judges.
- Each item set has the same mix of real items, controls, fillers, etc.
- There are standard methods – experimental psychology.

SEQUENCING THE ITEMS

- Random orderings.
- Explore all significant orderings.
- Maybe first few items are treated as “warm-up”
- There are standard methods – experimental psychology.

WHAT TO ASK THE JUDGES?

Various possibilities; e.g.:

“Is this a poem/story/joke/tune?”

“How good is this, on a scale of .....?”

“How original do you think this is, on a scale of .....?”

- Think carefully about the wording.
- Only a small number of questions.
- Keep them simple.

WHAT STATISTICS TO USE?

- Plan this before the testing.
- Have a hypothesis or two.
- What are you interested in:
  - Average ratings?
  - Best rating?
  - Spread of ratings?
  - Difference from control items?
- Lots of established practice in experimental psychology.

DON'T RUSH INTO IT

- Write up a detailed design first.
- Get comments on the design and revise it.
- Carry out a trial study ("pilot") to check your design.
- Fix any bugs found in the design
- Then do the main evaluation.

HOW TO EVALUATE THE QUALITY OF THE OUTPUT

Since “quality” is subjective, two main approaches:

- Rating of output items by naive judges.
- Rating of output items by experts.

A CASE STUDY : BINSTED'S JAPE SYSTEM

The JAPE program (Binsted 1996) produced "punning riddles": simple question-answer texts, supposedly humorous. E.g.

"What do you get when you cross a bird and a blunder? A fowl-up."

The program used a large dictionary, and 3 kinds of pattern-filling rules for 3 levels of the text generation.

JAPE EVALUATION (BINSTED 1996)

The data items

Texts generated by program in a structured sample.

Also control items of three types:

- human-written jokes
- randomly constructed texts
- sensible question-answer pairs.

All vocabulary and topics similar to the JAPE output.

JAPE EVALUATION (BINSTED 1996)

The presentation

Each sheet had 20 items (random order).

For each item, three questions:

- Was that a joke? [yes/no]
- How funny was it? [5-point scale]
- Have you heard it before? [yes/no]

Items also presented in spoken form through headphones.
No mention of “computer generation”.

JAPE EVALUATION (BINSTED 1996)

The experiment

Pilot study with 20 children; instructions amended as a result.

Main study: 122 children, (mostly) aged 8-11.

120 sheets properly completed, 2 spoiled.

All items rated by at least 9 participants (most items rated by 12).

JAPE EVALUATION (BINSTED 1996)

The analysis

“Was that a joke?”: Proportion of “yes” computed for each item.

“How funny was it?”: 1 – 5 . Mean computed for those items with “joke? = yes”.

Means computed for each of 4 types of item.

Compared means with Wilcoxon Signed Rank Test.

(Some exploratory retrospective adjustment of parameters.)

JAPE EVALUATION (BINSTED 1996)

The results

“Was that a joke?”: significant differences between:

- JAPE output & non-joke controls
- human jokes & non-joke controls
- human jokes & JAPE output

“How funny was it?”: same pattern of significant differences as above.

Human jokes **>** JAPE output **>** non-joke controls

HOW TO EVALUATE THE QUALITY OF THE OUTPUT

Since “quality” is subjective, two main approaches:

- Rating of output items by naive judges.
- Rating of output items by experts.

RATING BY EXPERT JUDGES

- Just a few judges.
- Judges should be knowledgeable.
- Questions can be more technical.
- Questions may need to be structured.

AGREEMENT BETWEEN JUDGES

Borrow methodology from existing areas; e.g.:

- ➢“annotation” - marking up text as a “gold standard”
- ➢“coding” – labelling data in the social sciences

“**reliability**” – to what extent do judges agree?

MEASURING AGREEMENT

Simpler if just two judges.

Simplest measure: the percentage of items which are labelled the same for all the judges.

But what about randomness -- might the judges agree (slightly) just by chance?

SOME AGREEMENT MEASURES

Cohen's Kappa (1960)
Two judges, adjusted for chance. Some weaknesses.

Cohen's Kappa (1968)
Like 1960, but with weightings for disagreements.

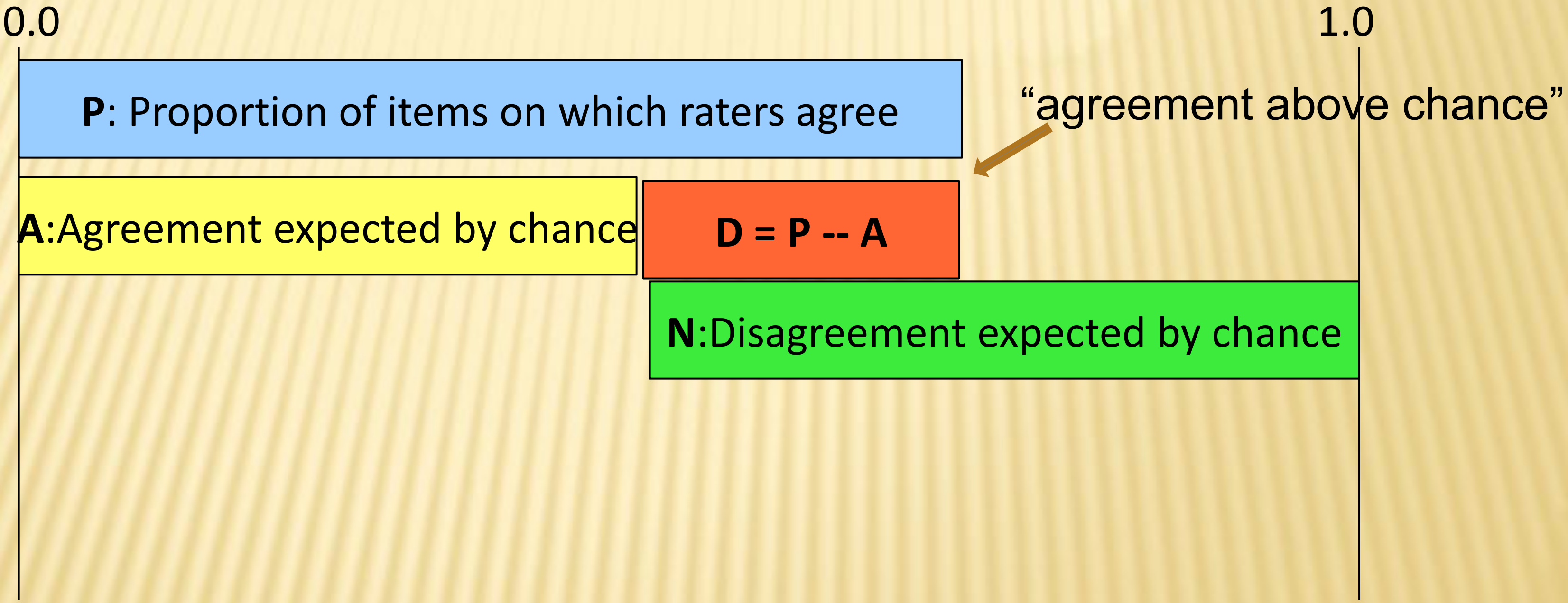
Scott's Pi
Two judges, adjusted for chance.

Fleiss's Kappa
Multiple judges, adjusted for chance.

Krippendorf's Alpha
Very general, covers many situations.

MEASURES OF AGREEMENT

The *Kappa* and *Pi* measures work this way:



0.0

1.0

P: Proportion of items on which raters agree

“agreement above chance”

A:Agreement expected by chance

D = P -- A

N:Disagreement expected by chance

Agreement score = (D/N)

MEASURES OF AGREEMENT

-1 (total disagreement)
+1 (total agreement)
0 (exactly chance-level agreement)

Informal guidelines for extent of agreement (for Cohen's 1960 Kappa):

- 0.0 < K <= 0.2 : slight
- 0.2 < K <= 0.4 : fair
- 0.4 < K <= 0.6 : moderate
- 0.6 < K < = 0.8 : substantial
- 0.8 < K : almost perfect (Landis & Koch 1977)

WARNING: *no consensus about such guidelines!*

SELECTION

- Some items may have no agreement.
- Which of the items will you use to test your hypotheses?

e.g. all those where more than half of the judges agree.

PART III : Other aspects

- Can it be automated?
- Amazon’s Mechanical Turk
- Non-controlled testing?
- Novelty
- Creativity

AUTOMATED TESTING?

In other fields, there are test suites, benchmark data, etc.

But for creative systems, quality is open-ended.

If there were mechanisable criteria – why not build them into the creative model?

AMAZON'S MECHANICAL TURK

https://www.mturk.com/mturk/welcome

Looks like an easy way to get judges.

But :

- very little information about the participants
- hard to check suitability of participants (knowledge, language, etc.)
- doubts about the motivation/conscientiousness of participants
- some practical constraints on fitting into the Amazon framework

Novelty

Distinguish (Boden 1990):

H-creativity: novel within history

P-creativity: novel for the creator

Correspondingly, two notions of “novelty”:

H-novelty
P-novelty

We are interested in P-creativity....
and hence **P-novelty.**

WHAT ABOUT NOVELTY?

Can we measure novelty in a similar way?

What questions should be asked?

Who would be suitable judges?

How can we detect P-novelty as opposed to H-novelty?

COULD THE SETTING BE MORE NATURAL?

E.g.

visual art in a gallery
music at a concert

How would we measure success?

A possible approach: ethnographic-style studies of the audience?? (cf. Shneiderman & Plaisant 2006)

CAN WE MEASURE EFFECTS?

Elsewhere, **task-based** evaluation is possible.

For creative systems, what is the "task"?

Perhaps "artistic" artefacts should evoke an **emotional response**.

For example – jokes: amusement.

What about forms that can evoke a wide range of emotions?
E.g. music, poetry, visual art, stories.

Could we measure this?

WHAT ABOUT CREATIVITY?

Can we measure creativity in a similar way?

What questions should the judges be asked?

Are we interested in a holistic judgement of “creative”?

Is there a risk of the word “creative” having different meanings for judges?

Who would be suitable judges?

WHICH METHODOLOGIES?

Is software engineering methodology irrelevant ?

No!

We still nee

- design pro
- implement

Still need abstraction, modularity, clear coding, etc.

Issues about requirement specification and testing are very different.

SUMMARY

- Evaluating a creative system is not like testing conventional software.
- If testing the “quality” of output, much of the methodology can be borrowed from the social sciences.
- There are still open questions about measurement of “novelty” or “creativity”.

“Empirical Methods for Artificial Intelligence”
Paul R. Cohen
MIT Press,Cambridge, Mass.
1995.