Computational Design Creativity Evaluation

David C. Brown
Worcester Polytechnic Institute, Worcester, MA, USA

This paper presents a simple framework for computational design creativity evaluation, presenting its components with rationale. Components are linked to recent computational creativity research in both art and design. The framework assumes that the product, not the process, is being evaluated, and that evaluation is done by comparison with descriptions of existing products using a set of aspects that each suggest creativity. Not every evaluation will use all of the components of the framework. It can be used to guide or assess design creativity research.

Introduction

This paper is concerned with the Computational Design Creativity (CDC) of engineered products and, specifically, the evaluation of creativity. That is, how does a computer system know when it "sees" a product that people will tend to label it as creative? A key issue is what the appropriate evaluation methods and measures are. Another is to identify the possible evaluators. Yet another is to describe what their knowledge might be.

In general, the issue is to determine all the types of ingredients involved in such an evaluation: hence the development of the ideas in this paper. The design creativity evaluation framework presented here consists of components, but not every evaluation will use all of the components.

There is no such thing *a priori* as a "creative computational design system", only one that produces artifacts that are evaluated as creative. This suggests that any CDC system must 'design *with* evaluation' and 'design *for* evaluation'. That is why this topic is so important.

Products, design descriptions, and design processes, are labeled as creative based on evaluation [1] [2] [3]. This framework is *not* concerned with processes, but it is possible that it might apply to them.

Design Computing and Cognition DCC'14. J.S. Gero (ed), pp. xx-yy.

The main assumptions are that evaluation is done *by comparison with descriptions of past or existing designs/products*, using a set of evaluative *aspects*, such as 'novelty', where each aspect may suggest creativity.

At this point it is still safe to say that humans are better creativity evaluators than machines [3] [4], and that (as with much of AI) the best initial approach to full computational evaluation of designs is to firmly base it on whatever we can determine about the way that humans do evaluation.

Figure 1 outlines the participants and their roles in the evaluation process. The design is assumed to be a description, while the product is a thing. The design might also be rendered virtually, and then evaluated. Despite being drawn with faces, some of these evaluations can be carried out by a CDC system: for partial designs and complete designs especially. An important issue is what each evaluator knows about the designer and vice versa.

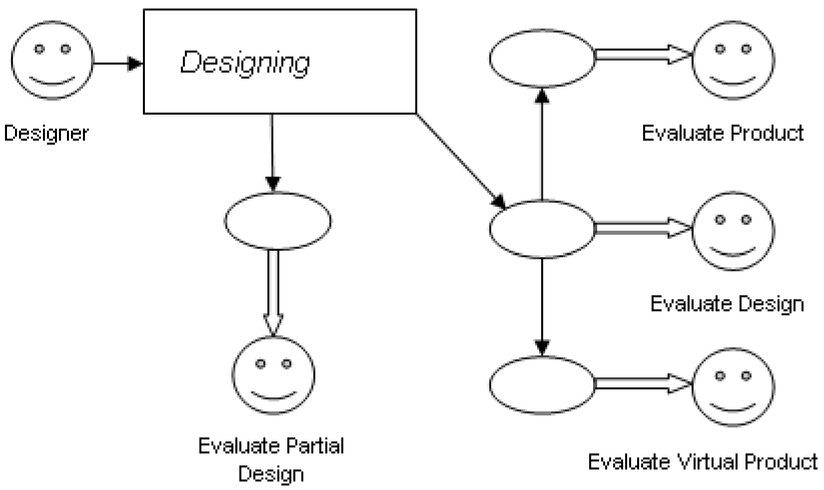

Fig 1. The participants in evaluation

There are many different factors that play a part in evaluation. For example, the time at which the evaluation is done is important for a CDC system. What varies then is how much of a design description is available. During designing it may be partial. After designing it should be complete. However, when presented with just the complete design description (or the actual product), the requirements may not be available, causing it to be much harder to evaluate relative to original intentions.

Evaluation of partial designs or of design decisions made during designing will need to be in terms of their *likely contribution* to the eventual perceived creativity of the final product. As this is difficult to predict, and such evaluation requires accurate expectations [5], partial designs are hard to evaluate. This may be made harder by the “new term” problem [6] where some previously unknown thing, property, or relationship is introduced during designing and must be recognized in order to make an effective evaluation.

Evaluation for creativity after the product has been designed is the norm. However, creativity evaluation of sub-parts and sub-systems during designing seems necessary in order to help drive the process towards a creative conclusion. Consequently, evaluations both *during* and *after* designing are needed for CDC systems.

The framework for evaluation that is proposed here is *simple* in that it provides a framework tuned to designing that has relatively few components, but it is *challenging* because to do computationally all that it suggests is currently very difficult. Galanter [7] states that “evaluation victories have been few and far between”. We expect it to remain difficult for quite a while. However, the framework should encourage researchers to try to implement all of its parts, rather than just a few. By specifying how each component of the framework is realized, it should allow researchers to classify how evaluation is done in existing and planned CDC systems. The framework addresses level 8, CC processes, in Sosa & Gero’s [8] Multi-level Computational Creativity model, with a nod towards levels 4 and 6, Product, and Cognition.

The references provided in this paper are a resource that should allow easy access to the current literature on design creativity evaluation, focusing primarily on the product, hardly at all on the process [2], and not at all on the designer’s personality [9] [10].

The field of Computational Creativity has probably advanced most in the area of the arts, with computer systems that paint, draw, write poetry, and interact with visitors/viewers. There appears to be very little reference by the design researchers to artistic creativity research, and vice versa. That provides additional motivation to do so here.

Computational artistic creativity researchers refer to “Aesthetic Evaluation”, as the arts are more concerned with beauty, and with taste [7]. The design area focuses on use and function, in addition to novelty. In the arts, novelty is a ‘given’ and function is usually secondary: however, that’s not to say that an artistic work has no function.

Romero et al. [11] argue that aesthetic judgment should apply to ‘form’ not ‘content’. So, for example, an artwork with the intended effect (function) of providing propaganda (content) should be judged aesthetically solely by the way it looks. However, this position needs to be softened for interactive digital art (see below). In contrast, for engineering design, usefulness and functionality are usually included in any evaluation.

We continue by sampling work on the evaluation of artistic creativity, in order to augment existing concepts from computational design creativity research. We then present the simple framework for design creativity evaluation, followed by an explanation of its components.

The Evaluation of Artistic Creativity

We proceed by reviewing *some* of the work in computational artistic creativity that relates to evaluation, evaluation knowledge, types of participants in the creative process, and creativity models. The hope is that we will find ideas to inform a framework for evaluation in design. Note that this is not intended to be a comprehensive review.

Evaluation Knowledge

Cohen et al. [6] presents a discussion of many aspects of evaluation in creativity. They introduce the idea of the evaluator's perspective or role, and the notion that what gets evaluated, and how, may change because of that role. They propose that a role may be "creator or designer, viewer, experiencer, or interactive participant".

They point out that much evaluation is concerned with "prediction": i.e., what impact a decision made during designing might make on the reaction to the final artwork, or how the *whole* artwork might be evaluated by others. Evaluation during designing is "directed to how to proceed".

Cohen et al. point out that prediction of an *emotional* response might be needed, not just an evaluation based on some "aesthetic principles" (what we call "aspects" in the framework below).

They also indicate the importance of the knowledge needed for evaluation: the artist's knowledge, knowledge about the artist, about cultural norms, the factors driving the creative act (which would include Requirements for designing), and the observer's knowledge. Note that the knowledge of the various evaluators may overlap but it is not likely to be the same. Some knowledge might be required to turn quantities (e.g., product dimensions) into qualities that can form part of an evaluation (e.g., stylish shape).They conclude that "Knowledge and experience emerge as decisive factors in producing artifacts of high creative value" (p. 98).

Evaluating Creativity

In Candy's "Evaluating Creativity" [12] she briefly discusses "Creativity in Design" but with almost no mention of the Engineering Design issues or the references introduced in this paper. Candy contrasts "digital arts" with Engineering Design with the former having the "designer and implementer" being the same person.

She uses the matrix for evaluating creativity proposed in Candy & Bilda [13] that is tuned to Art, and Interactive Digital Art in particular. Their model uses the well-known People, Process, Context (also known as "Press"), and Product divisions [14], with evaluation criteria added for each.

For Product evaluation she proposes Novel, Original, Appropriate, Useful, Surprising, Flexible, Fluent, and Engaging as the evaluation aspects. Interestingly, she also proposes measuring the *interaction* with the product with additional aspects: Immediate, Engaging, Purposeful, Enhancing, Exciting and Disturbing. Note the use of both functional and emotional terms. With less emphasis on interactive artwork, Candy & Bilda [13] also propose evaluating based on Composition, Aesthetic, Affect, Content and Technique. Note that here too the suggestion that an attempt be made to evaluate the emotional impact being made on the viewer/user. This is not just important for art, but for design too [15].

Candy also lists the people involved as artist/designer, participant/performer, and sometime "jury". For engineering design this reduces to designer and user, but rarely a jury.

Interaction with Art

Interactive art is often seen as providing "creative engagement" [13] by the viewers, or participants. In engineering design, there is much less concern about the creative nature of the 'use' of a product (i.e., whether it can be used or interacted with *creatively*).

However, the user or some other external evaluator will interact with a product in order to evaluate it, either by *viewing* it, *touching* it, *lifting* it, *manipulating* it, or *using* it for some task [16]. These interaction-based aspects are already well represented in most published sets of evaluation criteria (e.g., [17]): in fact without such interaction there can be limited evaluation of an implemented product. This is true even if the interactions are visualized or mentally simulated based on design descriptions or CAD models.

In design research we usually consider "a design" to be a description, and not the actual product. Descriptions do not usually exist as a deliverable in most artistic endeavors. For designing then we might evaluate at either the partial or full description stage, at the virtual implemented (e.g., CAD model) or the real implemented product stage. Hence we may need to be quite precise when talking about interaction enabling evaluation.

Candy & Bilda [13] have very distinct meanings for "interaction". They refer to "static", "dynamic-passive", "dynamic-interactive" and "dynamic-interactive varying" types of interaction with art. These correspond roughly to the "viewing" to "using" interactions mentioned above for designs.

They offer some guidelines for supporting/enhancing creative engagement with an interactive system. For example, "set expectations of the audience before they start to interact", suggests the possibility of improving eventual evaluation by invoking knowledge of familiar, existing products with the chosen structure, function or behavior of the new product. It also suggests taking advantage of perceived affordances or providing "signifiers" [18].

Computational Creativity Theory

Colton et al. [2] present a framework that is the basis of their Computational Creativity Theory, intended to "describe the processing and output of software designed to exhibit creative behaviour." Clearly this work is relevant for CDC systems even though there is some bias towards artistic creativity, as well as bias towards process and not just product. Note that their framework is not a framework about evaluation, and evaluation of the kind needed for product design is not stressed.

Colton et al. divide generative/creative "acts" into types g="ground" (producing new artifacts) and p="process" (producing new processes). These are coupled with what type of thing is being manipulated: expressions of concepts (E), concepts (C), aesthetic measures (A), and framing information (F): referred to as "FACE" tuples. This allows a rich set of creative actions to be described: in particular, and most original, actions that produce methods for generating concepts, or that produce methods for generating aesthetic measures. Actual things, such as concept descriptions, "expressions" (i.e., instances) of concepts, or measures, are seen as the input and output of these acts. Tuples of acts, such as $<A^g, C^g, E^g>$, together indicate a more complex creative act consisting of generative acts with information flowing between them.

From the examples provided it appears that evaluations apply to the whole new expressions of concepts (i.e., "designs" in our terms). If so, this misses the possibility that intermediate designs or design decisions might be evaluated.

With regard to evaluation, this paper suggests that there might be a minimal acceptable aesthetic (evaluation) level below which a result can be considered as “too low quality”. This seems very context-dependent: certainly in terms of the experience/knowledge of the evaluator. However, it does permit meta-level measures such as “average”, “best ever” and “precision”, which are an important way to evaluate a CDC system *over time*.

Colton et al. provide little evidence for which aesthetic measures may be appropriate (apart from novelty); nor how they might be combined. They instead propose that an “audience” judge both how much they “like” a creative act, and how much “cognitive effort” they were prepared to spend understanding that act.

This proposal is supposed to prevent an audience from having to “evaluate creativity directly”. However, by using a ‘profile’ or ‘fingerprint’ of evaluation aspects a CDC system could judge (“directly”) when a partial or complete design would be likely to be evaluated as creative. If necessary, it might even be able to calculate a creativity score based on the intended use of the design or the characteristics of the intended users. Note that use of evaluation aspects in our framework does not presuppose such a calculation by human or computer.

Colton et al. also propose the existence of a *distance measure* that can be applied to creative acts (including output), and a *similarity threshold* for distance, below which one act is deemed too similar to another, and therefore of less worth. An *upper threshold* can be used to determine whether two creative acts are similar enough to even be sensibly compared. These are useful concepts, but are probably aspect-specific, as different aspects will focus on different features and attribute of the design.

These thresholds do allow for some interesting hypotheses about the stages of development of a creative software system, as well as some general metrics that apply to groups of evaluators, such as judging whether a system’s creative act has an impact that provokes “shock”, provides “instant appeal”, or is prone to “triviality”.

Their paper argues for not using measures of the “value of solutions” (how well it solves a problem) in favor of using the “impact of creations”. The authors appear to be using a very specific meaning for “value” so this use is consistent with their proposals. However, for design solutions, how well the problem is solved can be determined relative to requirements and to actual usage scenarios. Not only can a design solve a problem (satisfy a need) it may also have perceived or real “value”. With regard to “impact” they also have a specific meaning in mind, referring more to the impact of creative acts, rather than products.

Colton et al. do not separate out the knowledge needed by a system for an artifact that will be evaluated by different types of evaluators. Having a model of the evaluator can change the action of a designer or design system. Similarly, an evaluator’s judgments will change depending on his/her/its model of the designer. Thus any framework of design evaluation needs to include these knowledge possibilities.

There is a rich history of considering types of knowledge and their roles during designing, such as the roles for knowledge in design reasoning [19], knowledge level descriptions of designing [20] [21], and types of knowledge during learning while designing [22].

Colton et al.'s work provides an excellent beginning to a theory of computational creativity, with strong bias towards creative processes, and some bias towards art. Even though it may provide a framework in which evaluation can occur, it is not a framework for evaluation itself. Their "creative acts" should provide a way to evaluate the development of creativity in a complete software system, which is their goal. For designs it seems self-evident that people can evaluate the creativity of a product without knowing anything about the design or manufacturing process. This is the normal situation for products, hence this paper's focus on the product.

Work on aesthetic evaluation in computational artistic creativity provides ideas about who might be evaluating, some particular views about how a user might interact with a "product", and suggestions about methods for evaluation. These inform the framework proposed in the next section.

A Creativity Evaluation Framework for CDC Systems

In this section we present a set of components involved in design evaluation, focusing on the actions, the knowledge needed, and the context for evaluation. In this framework we refer to "evaluator", considering it mainly to refer to a single evaluator that is not the designer, but in some circumstances they will be the same agent.

Note that it is assumed that evaluation is done by comparison with descriptions of past or existing designs/products: hence this does not appear as a component of the framework. We assume that appropriate design descriptions can be searched for, found, organized, selected, or recreated when needed. Apart from suggesting that the components given below influence this activity, no claims are being made about how this basis for comparison is actually produced. We assume that the description languages for the items in the basis are appropriate and comparable.

The Framework

The proposed framework for creativity evaluation for CDC systems has the following components:

1. *a description of the complete or partial artifact being judged, and/or the actual artifact;*
2. *the agent judging (i.e., person, computer program, or group);*
3. *the temporal basis for comparison (e.g., the point in time or the period);*
4. *the source of the design basis for comparison (e.g., personal, group, industry, global);*
5. *the set of "aspects" to include in the evaluation (e.g., novelty, surprise, style, utility, etc.);*
6. *the method of evaluation for each aspect;*
7. *the method used to combine the evaluations of the aspects (if one exists);*
8. *domain knowledge used by the evaluator (i.e., their amount of domain expertise);*
9. *the evaluator's knowledge about the designer (e.g., performance norms for the designer's level of expertise);*
10. *knowledge about the audience at whom the evaluation is aimed;*
11. *knowledge of the design requirements;*
12. *knowledge of resource constraints (e.g., materials, or available design time);*
13. *the evaluator's knowledge of the artifact due to the type and duration of experience with it;*
14. *the evaluator's knowledge of the design process;*
15. *the emotional impact of the design on the evaluator;*
16. *other contextual factors that may have an impact (e.g., culture).*

An Explanation of the Framework

Creativity evaluation depends on the components listed above. We will add some explanation about each one in turn. No detailed consideration will be given here as to how easily each might be adopted, adapted and implemented for CDC system use. The author is fairly convinced that they all could be implemented, with varying degrees of ease and precision.

Clearly, not every component of the model needs to be included in every CDC evaluation, and not every attribute of an artifact needs to be included in an evaluation.

A description of the complete or partial artifact being judged, and/or the actual artifact:

The evaluator will judge a design or a partial design. A CDC system deals with descriptions, although it is possible that, in the future, CDC systems might be 'grounded' by visual and tactile ability that could be applied to (perhaps computer generated) prototype artifacts. Humans are more likely to deal with artifacts, but can also judge descriptions. For complete evaluation it is necessary to have multi-level descriptions (e.g., showing subsystems), and descriptions in terms of Function, Behavior and Structure (see Erden et al. [23] for a review).

Some work on creativity evaluation considers a *set* of designs from a single designer (e.g., in response to the same requirements). However, even though the judgment is about the set, the essence of this approach is still comparing a single design against others.

The agent judging:

A 'judge' of some sort evaluates a design for creativity: that could be a person, a group, or a computer program. A CDC system might have knowledge and reasoning based on any of these. In a multi-agent design system, for example, both the designer and the judge might be computer programs.

The temporal basis for comparison:

The temporal basis is a point in time, or a period, on which to base the samples of related objects, prototypes, or standards [24] that are used for comparison with the design being judged [25].

The judgment of creativity is a moving target, as any new artifact could be added to the basis for comparison, which changes any subsequent judgment of the same (or similar) artifact. Of course, that depends on the judging agent having access to the modified basis [26]. Note that any changes to the basis (e.g., its organization) due to the addition of a new design may have meaning, such as indicating novelty [27].

Creativity evaluation is always a judgment at a time. It can be, and usually is, set to "now", but it could be set in the past, yielding a hypothetical evaluation about whether an artifact might have been seen as creative at some past time. For a CDC system we're considering "now" to be at the time of designing. By setting both the temporal and the source bases appropriately, evaluations of "rediscoveries" can be made [28].

The basis is often sourced from a time period. The normal period tends to be the maximal one of all history: at least back to the point where the technology makes comparisons irrelevant (e.g., laser cutters compared to flint knives) (*cf.* Colton et al.'s "upper threshold"). The temporal basis can be especially important for evaluating novelty [29].

The source of the design basis for comparison:

This component refers to from where the design basis is gathered. It might be strictly personal; in which case the basis is only designs produced by the designer (see [30]). This corresponds to evaluating for Boden's P-Creative designs, where P stands for Psychological [1]. By widening it to a group, industry, or global, and by using "all history" as the temporal basis, we are evaluating for H-Creative designs, where H stands for Historical.

This makes it clear that P- and H-creativity are labels for very particular areas of the *time-and-source space* of possible bases for comparison: i.e., just referring to P-Creative and H-Creative is much too simple.

As already mentioned, the actual basis for comparison is not considered to be part of this framework as it is considered to be 'generated by selection' depending on the time-and-source space specified.

In contrast to the evaluation of a single design against past designs, which might be called "absolute" creativity, some researchers evaluate a design, or a set of designs, against designs produced (often at the same point in time, and from the same requirements) from other designers in the same cohort [31] [32] [33]. This is often associated with the evaluation aspects of quantity and variety of the ideas generated. This limited comparison might be called "relative" creativity. However, both types can be accounted for by using the time and source components in this framework.

The set of "aspects" to include in the evaluation:

There are a very wide variety of different aspects mentioned in the literature that might be included for creativity evaluation, such as novelty, surprise, style, functionality, and value [29] [32] [33] [34] [35] [36] [37] [38] [39] [40] [41]. The field of artistic creativity evaluation has alternative (but overlapping) sets of aspects.

Besemer [17] has one of the most long-lived (from 1981) and well tested lists of aspects organized into categories. She includes Novelty (Surprising, Original), Resolution (Logical, Useful, Valuable, Understandable), and Style (Organic, Well-crafted, and Elegant).

Cropley & Kaufman [42] go even further, proposing 30 indicators of creativity that they experimentally reduced to 24. Their categories of aspects include Relevance & Effectiveness (Performance, Appropriateness, Correctness), Problematization (Prescription, Prognosis, Diagnosis), Propulsion (Redefinition, Reinitiation, Generation, Redirection, Combination), Elegance (Pleasingness, Completeness, Sustainability, Gracefulness, Convincingness, Harmoniousness, Safety), and Genesis (Vision, Transferability, Seminality, Pathfinding, Germinality, Foundationality).

The method of evaluation for each aspect:

Whichever aspects are included in a CDC system, an actual evaluation needs to be made using those aspects [43]. For example, an artifact needs to be judged for its novelty/originality [29] [38] [40] [44] [45] or for whether it is surprising [46] [47]. Different evaluation methods are possible for both of these aspects.

For example, novelty can be evaluated using a frequency-based approach that detects how many other designers have produced a similar design: the fewer the better. Novelty can also be estimated by accumulating the distance between the new design and the most similar design(s). If past designs are clustered, with some stereotypical design representing each cluster, the distance between the new design and the closest stereotype might also be used to evaluate novelty. Alternatively, if the new design causes re-clustering then this might indicate novelty. Finally, novelty might be measured by the amount of variation from the path of changes to features that designs with this functionality have exhibited over time: large variation suggests novelty. We conjecture that different methods will also exist for other aspects besides novelty.

In addition, depending on the design description used, it may be possible to apply the evaluation of aspects to different levels of abstraction in the description [32] [48] [49], and to descriptions that include Function, Behavior and Structure [40].

The method used to combine the evaluations of the aspects:

Overall evaluations have strengths; therefore artifacts may be seen as more, or less, creative – i.e., it isn't a Boolean decision. However, if many aspects are evaluated this will produce a 'profile' of the amount of creativity demonstrated across all those aspects, not a single result [17]. Evaluation in a single, combined dimension results from the evaluator's biases about how to combine different aspect evaluations [31] [32] [37] [40].

Even if a particular evaluating agent is being modeled (e.g., an actual user or group of users), this combination method may not exist explicitly. Evolutionary methods have been used to produce combinations of aspects with some success [7] but often the methods of combination they produce "seem alien". Learning systems exist that extract and use features to do "aesthetics-based evaluation" [11] [50]. Fuge et al. [51] describe a method that is able to learn to mimic expert creativity ratings, such as “variety” scores.

A complex issue regarding combining evaluations that needs addressing is how the separate evaluations of creativity in the Function, Behavior and Structure levels affect each other and the evaluation of the whole artifact. For example, a candle that produces sparks on the hour to indicate time provides a standard function by behaving in a novel way, with only a slightly new structure: how creative is that?

The domain knowledge used by the evaluator:

It is well established in the literature (see [43]) that the amount of domain expertise that the designer has makes a big difference to their potential for creativity. However, to *fully* appreciate a design the evaluator needs to (at least) match their level of sophistication. For example, expert evaluators may know about complex electromechanical devices: less expert designers may only know about Legos. Hence the nature and amount of the evaluator’s domain knowledge will make a big difference to the evaluation [42]. Note that this need not be put explicitly into a CDC system – in fact it may not be able to be – but it might be accumulated using machine learning.

The knowledge about the designer:

Knowledge of the capabilities of the designer may play a role in creativity evaluation: for example, the evaluator might be able to recognize Transformational creativity [1] [52]. Also, knowing the performance norms for the designer’s level of expertise is important. Consider a design description of a building from a 10 year old child versus a design description from an excellent Architect. An excellent child might be very creative relative to what they’ve already done (P-Creative), while an excellent architect is more likely to be judged as very creative relative to what everyone else has already done (H-Creative).

The knowledge about the audience at whom the evaluation is aimed:

The evaluation must be understandable by the recipient of the evaluation. What you’d tell a child would be different from what you’d tell an expert. The conjecture is that this is not just a matter of the type of language used for the evaluation report, but that the actual evaluation might vary. For example, if a simple Yes/No or numeric position on a scale answer is desired then a powerful general technique such as CSPs, Neural Nets, or Evolutionary Computing might be used for the evaluation, as rationale for either the design or the evaluation is not needed, nor available. If the evaluation is for an expert, then it might be provided in technical terms, and mention product features, for example: whereas an evaluation of a process for an expert might mention ingredients such as selection, planning, evaluation, constraint testing, patching, failure handling, etc.

The knowledge of the design requirements:

Do the 'requirements' for the product, possibly including the intended function, need to be known to evaluate creativity? We argue that it is not necessary, but it should be helpful, as it allows the basis for comparison to be more precisely selected.

The knowledge of resource constraints:

If an evaluator understands how a designer dealt with resources constraints, such as limits on material availability or limited design time, it can affect their creativity evaluation.

The evaluator's knowledge of the artifact due to the type and duration of experience with it:

An evaluator might read the design description, see the artifact, touch the artifact, manipulate the artifact, or actually use the artifact [16]. This affects the completeness of their understanding of the artifact, and therefore their evaluation. Ludden's example involves 'surprise', but other aspects could also be affected.

Cohen et al. [6] conjecture about the computer 'experiencing' a design, and whether perception is required in order to do evaluation that matches what humans do. A CDC will need to have the equivalent of the ability to 'imagine' a design when given a design description, in order to evaluate its look or feel, its organic qualities, or its use.

The evaluator's knowledge of the design process:

Colton [53] argues that, especially for artistic products, knowledge of the process is extremely important for the evaluation of creativity. However, it is clear that a very novel and interesting process might result in a not very creative design. We include this component of the framework for completeness. For many of the researchers referenced in this paper, this component is not essential for the evaluation of designed artifacts.

The emotional impact of the design on the evaluator:

There is an increasing amount of interest in the emotional impact of designs [54]. But what is the role of emotion in the evaluation of creativity? The "impact" on the evaluator does play a role [34] but how do fun, cuteness, cleverness, memories, or jokes play a role in evaluation? Horn & Salvendy [36] claim that arousal and pleasure influence the evaluation of product creativity. Cropley [55] points out that "departure from the usual arouses discomfort" and perhaps "departure from the usual arouses excitement". Datta et al. [50] relate emotional impact to aesthetics evaluation.

Norman [15] proposes that initial design evaluation takes place at a sensory/visceral level, where 'appearance' can evoke an emotional response. The behavioral level of evaluation is concerned with usability: a very good or very bad experience can evoke corresponding emotions (e.g., frustration). The reflective level of evaluation is about prestige and desirability: i.e., how having the product makes one feel, and the degree of good taste that it might convey. It's clear that some of the aspects introduced above (e.g., Besemer's "Style" dimension) might act as a proxy for some of the emotional response, while the prestige associated with a particular product or designer could be estimated.

In general, emotional impact is clearly a difficult component to include in a CDC system. However, it might be detected or estimated in a variety of ways: *direct* methods such as eye movement/dilation, galvanic skin response, and brain wave changes; *indirect* methods such as measures of similarity to products that have known emotional impact, or classifiers trained using machine learning from user reporting.

Other contextual factors that may have an impact:

This, we must admit, is a catch-all category. However, there are factors, such as culture, that may play a role in evaluation that could go here, as it isn't clear that they always apply or are a main influence for CDC systems.

One such factor is whether a past artifact has been acknowledged as creative: perhaps to the point of it being a disruptive product, changing the direction of future artifacts in the same category. Sternberg et al. [28] describe this as "propelling" a field. This knowledge might be used to suggest that a new artifact might be creative by analogy: if the new artifact (X) has 'similar' characteristics to an existing artifact (Y), and Y was seen as creative and influential in the past, then perhaps X will be seen as a creative influence. Such an evaluation would be helped by having similarity information [56] available, and knowledge about the design time. Of course, too much similarity decreases novelty.

Some evaluation schemes include "usefulness" and the "importance" of the use as evaluation aspects (e.g., [40]). This might be measured in terms of actual use, or potential use. As the artifact has just been, or is still being, designed, evaluating "actual" use will not be possible. There needs to be enough knowledge included during the design creativity evaluation process to estimate how much it might be used, and weight it by "importance" or potential "impact".

Summary & Conclusion

The framework presented here differs from other work by focusing on artifact design, the different types of participants in the evaluation process, and the types of knowledge needed by the designer and the evaluator: in particular what each needs to know about the other.

This framework is intended to be used to guide or assess design creativity research, with the hope that it will eventually apply to CDC systems. The references should allow easy access to the current literature on design creativity evaluation. Given the number and difficulty of the components in the framework it is obvious that CDC systems still need a lot of work. The framework also makes it clear that how creative an artifact is may only be properly stated if the full context of the evaluation is included.

References

1. Boden MA (1994) What is Creativity? Dimensions of Creativity, M.A. Boden (Ed.), The MIT Press, 75-117.
2. Colton S, Charnley J, Pease A (2011) Computational Creativity Theory: The FACE and IDEA Descriptive Models. Proc. 2nd Int. Conf. on Computational Creativity, 90-95.
3. Hennessey BA, Amabile TA (2010) Creativity. Ann. Rev. Psychology 61: 569-598.
4. Amabile TM (1996) Creativity in Context. Westview Press.
5. Grecu DL, Brown DC (2000) Expectation Formation in Multi-Agent Design Systems. Proc. Artificial Intelligence in Design'00, J.S. Gero (Ed.), Kluwer, 651-671.
6. Cohen H, Nake F, Brown DC, Brown P, Galanter P, McCormack J, d'Inverno M (2012) Evaluation of Creative Aesthetics. Computers and Creativity, J. McCormack & M. d'Inverno (Eds.), Springer-Verlag, 95-111.

7. Galanter P (2012) Computational Aesthetic Evaluation: Past and Future. Computers and Creativity, J. McCormack & M. d'Inverno (Eds.), Springer-Verlag, 255-293.
8. Sosa R, Gero JS (2013) Multilevel Computational Creativity. Proc. 4th Int. Conf. on Computational Creativity, 198-204.
9. Eysenck HJ (1994) The Measurement of Creativity. Dimensions of Creativity, M.A. Boden (Ed.), The MIT Press, 199-242.
10. Charyton C, Jagacinski RJ, Merrill JA (2008) CEDA: A Research Instrument for Creative Engineering Design Assessment. Psychology of Aesthetics, Creativity, and the Arts 2(3):147-154.
11. Romero J, Machado P, Carballal A, Correia J (2012) Computing Aesthetics with Image Judgment Systems. Computers and Creativity, J. McCormack & M. d'Inverno (Eds.), Springer-Verlag, 295-321.
12. Candy L (2013) Evaluating Creativity. Creativity and Rationale: Enhancing Human Experience by Design, J.M. Carroll (Ed.), Springer-Verlag, 57-84.
13. Candy L, Bilda Z (2009) Understanding and evaluating creativity. Creativity & Cognition, 497-498
14. Rhodes M (1961). An analysis of creativity. Phi Delta Kappan, 42:305-310.
15. Norman DA (2004) Emotional Design: Why We Love (or Hate) Everyday Things. Basic Books.
16. Ludden GDS, Schifferstein HNJ, Hekkert P (2008) Surprise as a Design Strategy. Design Issues 24(2):28-38.
17. Besemer SP (2006) Creating products in the age of design. How to improve your new product ideas! New Forums Press, Inc.
18. Norman DA (2008) THE WAY I SEE IT: Signifiers, not affordances. Interactions Magazine, ACM, 15(6):18-19.
19. Brown DC (1992) Design. Encyclopedia of Artificial Intelligence, 2nd edition, S.C. Shapiro (Ed.), J. Wiley.
20. Smithers T (1996) On knowledge level theories of design process. Artificial Intelligence in Design'96, J.S. Gero & F. Sudweeks (Eds.), Kluwer, 561-579.
21. Smithers T (1998) Towards a Knowledge Level Theory of Design Process. Artificial Intelligence in Design'98, J.S. Gero & F. Sudweeks (Eds.), Springer, 3-21.
22. Sim SK, Duffy AHB (2004). Knowledge transformers: A link between learning and creativity. AIEDAM 18(3):271-279.
23. Erden MS, Komoto H, van Beek TJ, D'Amelio V, Echavarria E, Tomiyama T (2008) A Review of Function Modeling: Approaches and Applications. Special issue on Multi-modal Design, A. Goel, R. Davis & J.S. Gero (Eds.), AIEDAM 22(2).

24. Redelinghuys C (2000) Proposed criteria for the detection of invention in engineering design. Jnl. Engineering Design 11(3):265-282
25. Wiggins GA (2006) A Preliminary Framework for Description, Analysis and Comparison of Creative Systems. Jnl. of Knowledge Based Systems 19(7):449-458.
26. Sosa R, Gero JS (2005) A computational study of creativity in design: the role of society. AIEDAM 19(4):229-244.
27. Maher ML, Brady K, Fisher D (2013) Computational Models of Surprise in Evaluating Creative Design. Proc. 4th Int. Conf. on Computational Creativity, University of Sydney, 147-151.
28. Sternberg RJ, Kaufman JC, Pretz JE (2002) The Creativity Conundrum: A propulsion model of kinds of creative contributions. Psychology Press.
29. Maher ML, Fisher DH (2012) Using AI to Evaluate Creative Designs. Proc. 2nd Int. Conf. on Design Creativity (ICDC2012), 45-54.
30. Jagtap S, Larson A, Hiort V, Olander E, Warell A (2012) Ideation Metrics: Interdependency between Average Novelty and Variety. Int. Design Conf., DESIGN 2012, 1881-1892.
31. Oman SK, Tumer IY, Wood K, Seepersad C (2013) A comparison of creativity and innovation metrics and sample validation through in-class design projects. Research in Engineering Design 24(1):65-92.
32. Shah JJ, Vargas Hernandez N, Smith SM (2003) Metrics for measuring ideation effectiveness. Design Studies 24:111-134.
33. Kudrowitz BM, Wallace DR (2012) Assessing the Quality of Ideas from Prolific, Early-Stage Product Ideation. Jnl. Engineering Design, Special Issue on Design Creativity, (ifirst online preview), 1-20.
34. Christiaans HHCM (1992) Creativity in Design: The role of domain knowledge in designing. Uitgeverij Lemma BV.
35. Dean DL, Hender JM, Rodgers TL, Santanen EL (2006) Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation. Jnl. Association for Information Systems 7(10).
36. Horn D, Salvendy G (2006) Product creativity: conceptual model, measurement and characteristics. Theoretical Issues in Ergonomics Science 7(4):395-412
37. Ritchie G (2007) Some Empirical Criteria for Attributing Creativity to a Computer Program. Minds & Machines 17:67-99.
38. Srivathsavai R, Genco N, Hölttä-Otto K, Seepersad CC (2010) Study of Existing Metrics Used in Measurement of Ideation Effectiveness. Proc. ASME 2010 Int. IDETC & CIE Confs., DETC2010-28802.
39. Liikkanen LA, Hämäläinen MM, Häggman A, Björklund T, Koskinen MP (2011) Quantitative Evaluation of the Effectiveness of Idea Gen-

eration in the Wild. Human Centered Design, Lecture Notes in Computer Science, 6776, 120-129.
40. Sarkar P, Chakrabarti A (2011) Assessing design creativity. Design Studies 32:348-383
41. Lu C-C, Luh D-B (2012) A Comparison of Assessment Methods and Raters in Product Creativity. Creativity Research Jnl. 24(4):331-337.
42. Cropley DH, Kaufman JC (2012) Measuring Functional Creativity: Non-Expert Raters and the Creative Solution Diagnosis Scale. Journal of Creative Behavior, 46(2):119-137.
43. Brown DC (2013) Guiding Computational Design Creativity Research. Int. Jnl. of Design Creativity and Innovation 1(1).
44. Lopez-Mesa B, Vidal R (2006) Novelty Metrics in Engineering Design Experiments. Int. Design Conf., DESIGN 2006, 557-564.
45. Shelton KA, Arciszewski T (2007) Formal innovation criteria. Int. Jnl. Computer Applications in Technology 30(1/2):21-32.
46. Macedo L, Cardoso A, Reisenzein R, Lorini E, Castelfranchi C (2009) Artificial Surprise. Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence, J. Vallverdu & D. Casacuberta (Eds.), Information Science Reference (IGI Global).
47. Brown DC (2012) Creativity, Surprise & Design: An Introduction and Investigation. Proc. 2nd Int. Conf. on Design Creativity (ICDC2012).
48. Nelson BA, Yen J, Wilson JO, Rosen D (2009) Refined Metrics for Measuring Ideation Effectiveness. Design Studies 30:737-743.
49. Farzaneh HH, Kaiser MK, Schroer B, Srinivasan V, Lindemann U (2012) Evaluation of Creativity: Structuring Solution Ideas Communicated in Groups Performing Solution Search. Int. Design Conf., DESIGN 2012, 1871-1880.
50. Datta R, Joshi D, Li J, Wang JZ (2006) Studying Aesthetics in Photographic Images Using a Computational Approach. Lecture Notes in Computer Science, 3953, Proc. European Conf. Computer Vision, Part III, 288-301.
51. Fuge M, Stroud J, Agogino A (2013) Automatically Inferring Metrics for Design Creativity. ASME IDETC & CIE, DETC2013-12620.
52. Ritchie G (2006) The transformational creativity hypothesis. New Generation Computing 24:241-266.
53. Colton S (2008) Creativity versus the perception of creativity in computational systems. Proc. AAAI Spring Symp. on Creative Systems.
54. McDonagh D, Denton H, Chapman J (Eds.) (2009) Special issue on 'Design and Emotion'. Journal of Engineering Design 20(5).
55. Cropley DH (2009) Fostering and measuring creativity and innovation: individuals, organisations and products. Proc. Conf. Can Creativity be

Measured? Section 17, Education and Culture DG, European Commission, http://ec.europa.eu/education/lifelong-learning-policy/creativity-conference_en.htm
56. Minsky M (2006) The Emotion Machine: Commonsense Thinking, Artificial Intelligence & the Future of the Human Mind. Simon & Schuster.