Machine Learning for Design, Week 9

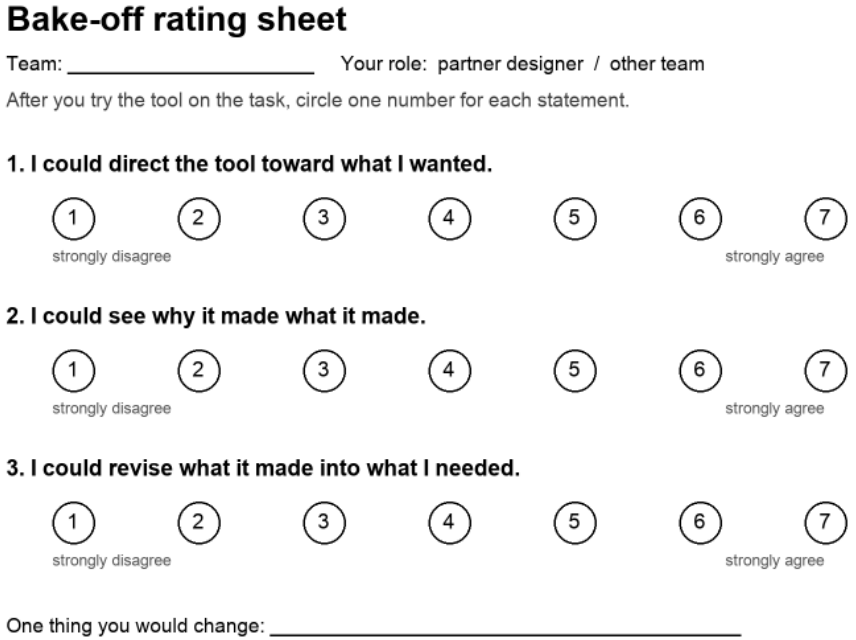
Evaluating Design Tools

Next week: the bake-off

Could I direct it?

Could I see why it made what it made?

Could I revise what it made?



Bake-off rating sheet

Team: _____ Your role: partner designer / other team

After you try the tool on the task, circle one number for each statement.

1. I could direct the tool toward what I wanted.

1 2 3 4 5 6 7

strongly disagree — strongly agree

2. I could see why it made what it made.

1 2 3 4 5 6 7

strongly disagree — strongly agree

3. I could revise what it made into what I needed.

1 2 3 4 5 6 7

strongly disagree — strongly agree

One thing you would change: _____

The rating sheet, with these three as statements

How do you show that a design tool helps designers?

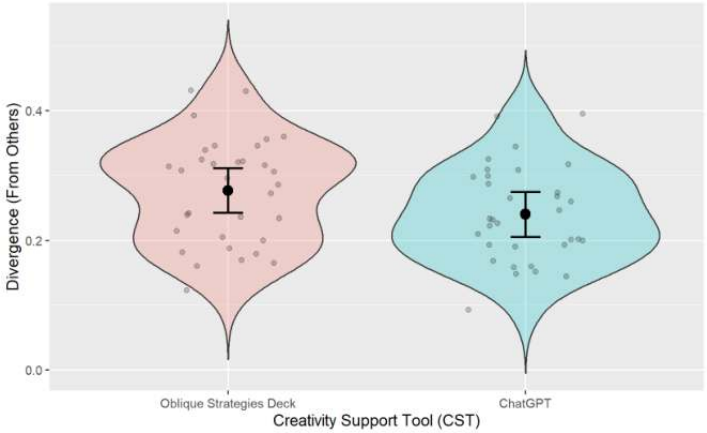
Efficiency and creativity

“tools that are not effective and efficient will probably hinder creativity, but it isn't clear that the reverse will hold”

Resnick et al., 2005

Week 1: ideas made with ChatGPT were closer to everyone else's

Designers using an AI image generator fixated more and had fewer ideas



How far each idea was from everyone else's, Oblique Strategies deck vs ChatGPT

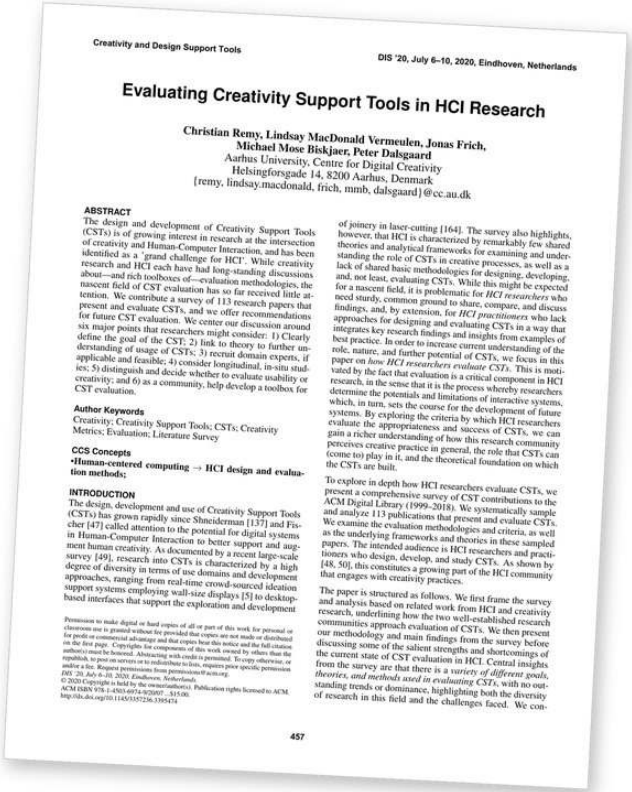
Resnick et al., 2005, Section 12; Anderson et al., C&C 2024, Fig. 2 (CC BY-NC-SA 4.0); Wadinambiarachchi et al., CHI 2024

Today

1. What to measure
2. The Creativity Support Index
3. Design principles
4. Discussion
5. Workshop: plan your bake-off test
6. Crit

1. What to measure

Optional reading: Remy et al., “Evaluating Creativity Support Tools in HCI Research,” DIS 2020



Evaluating Creativity Support Tools in HCI Research

Christian Remy, Lindsay MacDonald Vermeulen, Jonas Frich, Michael Mose Biskjaer, Peter Dalsgaard
Aarhus University, Centre for Digital Creativity
Helsingforsgade 14, 8200 Aarhus, Denmark
{remy, lindsay.macdonald, frich, mmb, dalsgaard}@cc.au.dk

ABSTRACT
The design and development of Creativity Support Tools (CSTs) is of growing interest in research at the intersection of creativity and Human-Computer Interaction, and has been identified as a 'grand challenge for HCI'. While creativity research and HCI each have had long-standing discussions about—and rich toolboxes of—evaluation methodologies, the nascent field of CST evaluation has so far received little attention. We contribute a survey of 113 research papers that present and evaluate CSTs, and we offer recommendations for future CST evaluation. We center our discussion around six major points that researchers might consider: 1) Clearly define the goal of the CST; 2) link to theory to further understanding of usage of CSTs; 3) recruit domain experts, if applicable and feasible; 4) consider longitudinal, in-situ studies; 5) distinguish and decide whether to evaluate usability or creativity; and 6) as a community, help develop a toolbox for CST evaluation.

Author Keywords
Creativity; Creativity Support Tools; CSTs; Creativity Metrics; Evaluation; Literature Survey

CCS Concepts
•**Human-centered computing** → **HCI design and evaluation methods;**

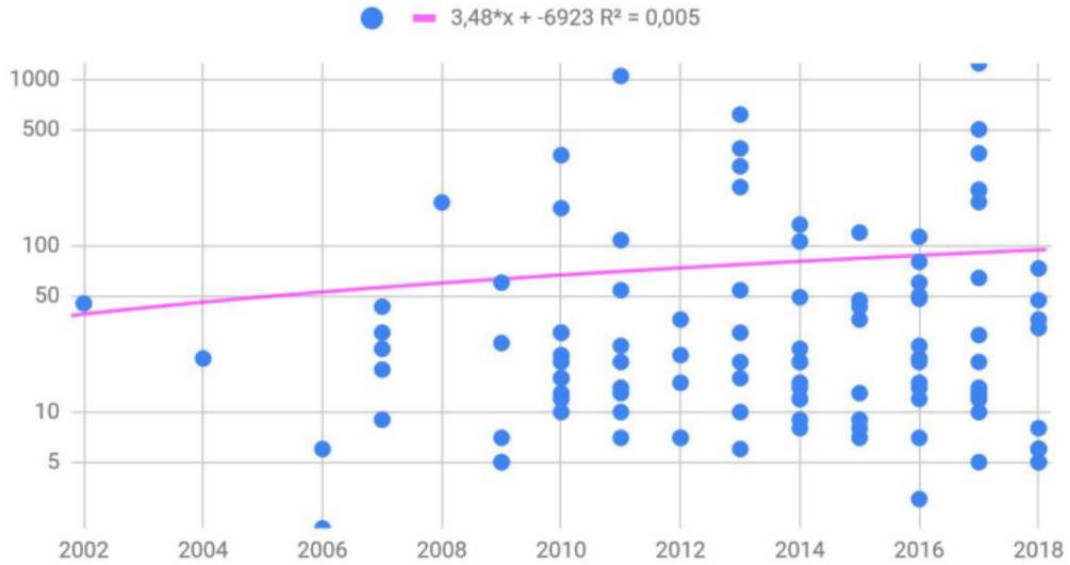
INTRODUCTION
The design, development and use of Creativity Support Tools (CSTs) has grown rapidly since Shneiderman [137] and Fischer [47] called attention to the potential for digital systems in Human-Computer Interaction to better support and augment human creativity. As documented by a recent large-scale survey [49], research into CSTs is characterized by a high degree of diversity in terms of use domains and development approaches, ranging from real-time crowd-sourced ideation support systems employing wall-size displays [5] to desktop-based interfaces that support the exploration and development of joinery in laser-cutting [164]. The survey also highlights, however, that HCI is characterized by remarkably few shared theories and analytical frameworks for examining and understanding the role of CSTs in creative processes, as well as a lack of shared basic methodologies for designing, developing, and, not least, evaluating CSTs. While this might be expected for a nascent field, it is problematic for *HCI researchers* who need sturdy, common ground to share, compare, and discuss findings, and, by extension, for *HCI practitioners* who lack approaches for designing and evaluating CSTs in a way that integrates key research findings and insights from examples of best practice. In order to increase current understanding of the role, nature, and further potential of CSTs, we focus in this paper on *how HCI researchers evaluate CSTs*. This is motivated by the fact that evaluation is a critical component in HCI research, in the sense that it is the process whereby researchers determine the potentials and limitations of interactive systems, which, in turn, sets the course for the development of future systems. By exploring the criteria by which HCI researchers evaluate the appropriateness and success of CSTs, we can gain a richer understanding of how this research community perceives creative practice in general, the role that CSTs can (come to) play in it, and the theoretical foundation on which the CSTs are built.

To explore in depth how HCI researchers evaluate CSTs, we present a comprehensive survey of CST contributions to the ACM Digital Library (1999–2018). We systematically sample and analyze 113 publications that present and evaluate CSTs. We examine the evaluation methodologies and criteria, as well as the underlying frameworks and theories in these sampled papers. The intended audience is HCI researchers and practitioners who design, develop, and study CSTs. As shown by [48, 50], this constitutes a growing part of the HCI community that engages with creativity practices.

The paper is structured as follows. We first frame the survey and analysis based on related work from HCI and creativity research, underlining how the two well-established research communities approach evaluation of CSTs. We then present our methodology and main findings from the survey before discussing some of the salient strengths and shortcomings of the current state of CST evaluation in HCI. Central insights from the survey are that there is a *variety of different goals, theories, and methods used in evaluating CSTs*, with no outstanding trends or dominance, highlighting both the diversity of research in this field and the challenges faced. We con-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
DIS '20, July 6–10, 2020, Eindhoven, Netherlands.
© 2020 Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6974-9/20/07…$15.00.
http://dx.doi.org/10.1145/3357236.3395474

Evaluation methods in 113 papers



Participants in each evaluation, by year (log scale)

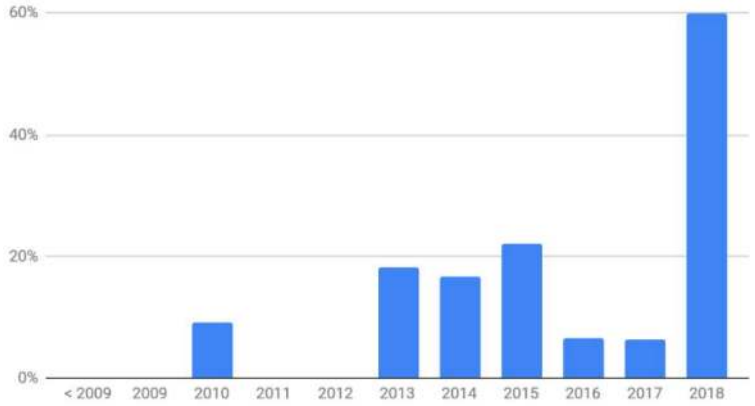
Most used methods:
observation (38)
surveys or questionnaires (37)
interviews (33)

Only **3** used the
Creativity Support Index

Remy et al., DIS 2020, Fig. 4 and Results

Six recommendations from 113 papers

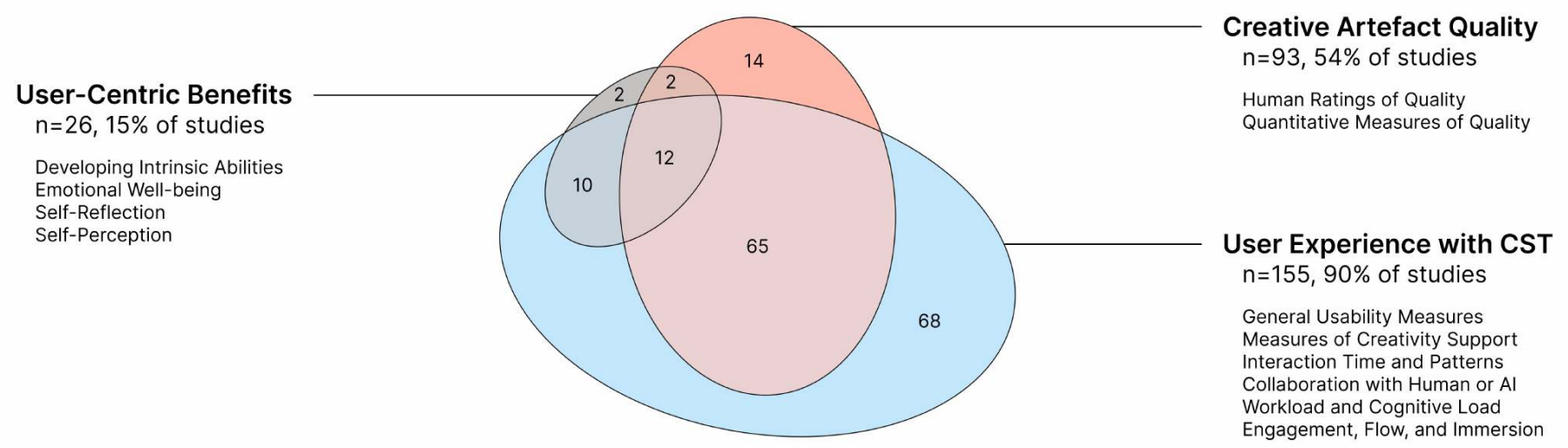
1. Define the tool's goal
2. Link to theory
3. Recruit domain experts
4. Study use over time, in place
5. Decide: usability or creativity
6. Share evaluation tools



Share of evaluations that were usability studies, by year

Remy et al., DIS 2020, abstract and Fig. 5

Outcome measures in 173 evaluations

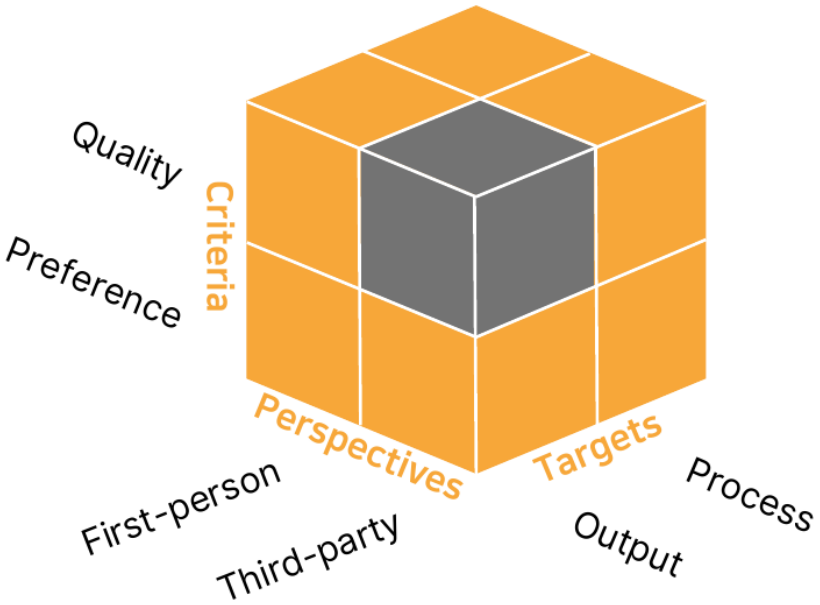


Few asked what the person gained, such as learning or well-being

They call for validated measures for generative AI tools

Cox et al., “Beyond Productivity: Rethinking the Impact of Creativity Support Tools,” C&C 2025, Fig. 1 (CC BY 4.0), optional reading

HALIE's three dimensions



Targets: process, not just output

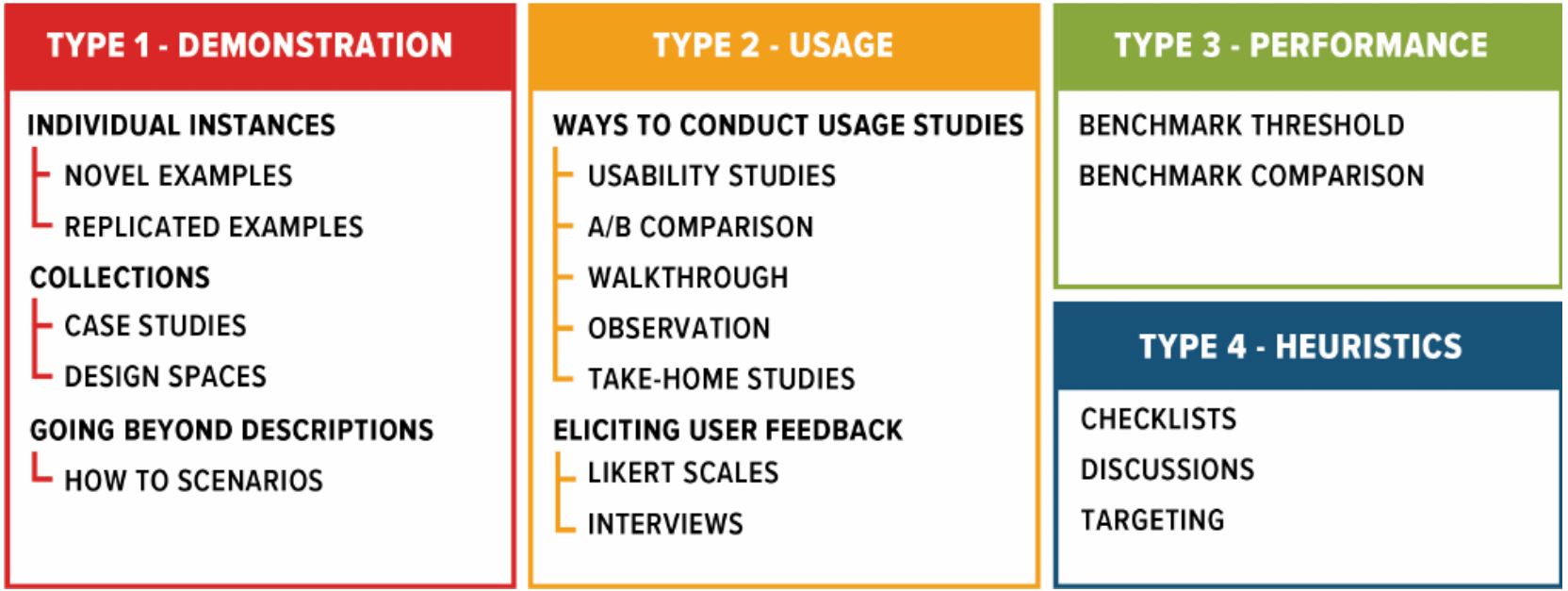
Perspectives: first-person, not just third-party

Criteria: preference, not just quality

The model that scored best without people was not always the best one to work with

Lee et al., “Evaluating Human-Language Model Interaction,” TMLR 2023, Fig. 1, optional reading

Evaluating a toolkit



66 of 68 toolkit papers used demonstrations, **19** of them alone

Ledo et al., CHI 2018, from 68 toolkit papers (optional reading)

Measuring our slider









prompt + simple	−1.5	−0.75	0	+0.75	+1.5	prompt + ornate
How ornate, by CLIP						
−0.075	−0.082	−0.082	−0.085	−0.055	−0.037	+0.055
Same chair as at 0, by DINOv2						
0.93	0.99	0.99	1.00	0.95	0.93	0.52
Seconds per image, on a laptop						
47	128	120	49	118	122	48

One prompt and seed, from week 5. CLIP ViT-B/32: similarity to the ornate words minus similarity to the simple words. DINOv2: similarity to the image at 0.

2. The Creativity Support Index

Today's reading: Cherry & Latulipe, “Quantifying the Creativity Support of Digital Tools through the Creativity Support Index,” TOCHI 2014



Quantifying the Creativity Support of Digital Tools through the Creativity Support Index

21

ERIN CHERRY and CELINE LATULIPE, University of North Carolina at Charlotte

Creativity support tools help people engage creatively with the world, but measuring how well a tool supports creativity is challenging since creativity is ill-defined. To this end, we developed the Creativity Support Index (CSI), which is a psychometric survey designed for evaluating the ability of a creativity support tool to assist a user engaged in creative work. The CSI measures six dimensions of creativity support: Exploration, Expressiveness, Immersion, Enjoyment, Results Worth Effort, and Collaboration. The CSI allows researchers to understand not just how well a tool supports creative work overall, but what aspects of creativity support may need attention. In this article, we present the CSI, along with scenarios for how it can be deployed in a variety of HCI research settings and how the CSI scores can help target design improvements. We also present the iterative, rigorous development and validation process used to create the CSI.

Categories and Subject Descriptors: H.5.2 [**User Interfaces**]: Evaluation/Methodology; Standardization

General Terms: Human Factors, Reliability, Standardization

Additional Key Words and Phrases: Creativity support tools, evaluation, psychometrics, surveys

ACM Reference Format:

Erin Cherry and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the creativity support index. ACM Trans. Comput.-Hum. Interact. 21, 4, Article 21 (May 2014), 25 pages.

DOI: http://dx.doi.org/10.1145/2617588

1. INTRODUCTION

Creativity is revered by society as an essential human resource: it is linked to the economic principles of innovation and growth, as well as to individual personal development through creative expression, intellectual challenge, and psychological flourishing [Amabile 1996; Csikszentmihalyi 1997]. Creativity support tools (CSTs) can make a substantial impact on both individuals and society by improving scientific, engineering, humanist, and artistic endeavors [Latulipe 2013; Shneiderman 2007]. CSTs are able to influence creative work that happens every day, in addition to helping to bring about those rare creative moments that can change history and have vast impacts on society. In addition to supporting creative work, CSTs also have the potential to help people on their own personal creative journeys. However, it is challenging to evaluate the ability of CSTs to support or influence creative work, because creativity itself is not easily defined and measured [Beghetto 2007; Kaufman and Beghetto 2009].

Our contribution in this article is a quantitative, psychometric tool, called the Creativity Support Index (CSI), which is a survey to assess a tool's creativity support.

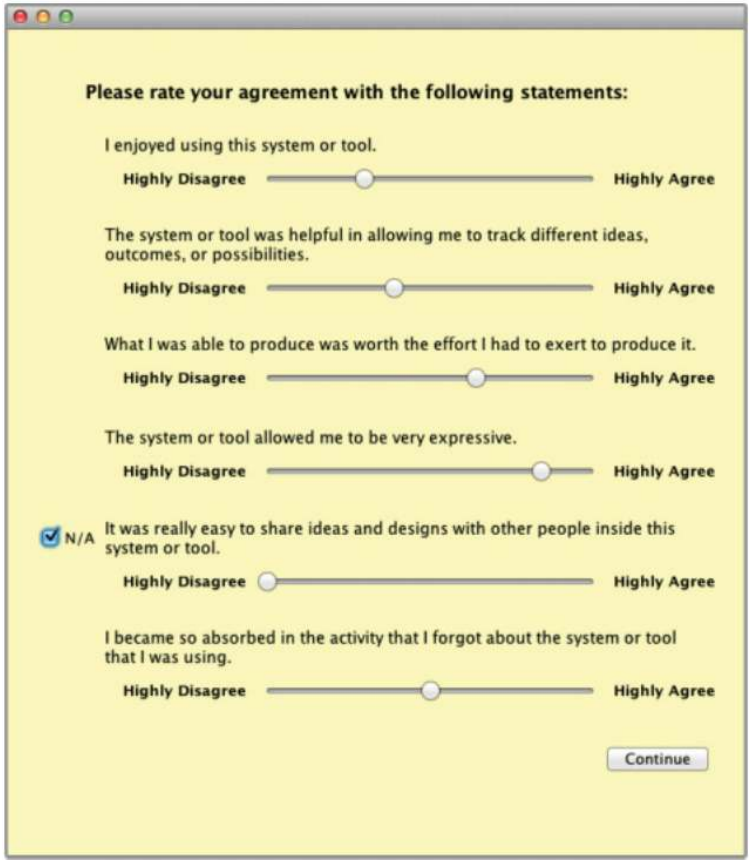
This work is supported by the National Science Foundation Creative IT program, under grant IIS-0855882.

Authors' address: Erin Cherry (formally Carroll), University of Rochester, The Computer Science Dept., PO Box 270226, Rochester, NY 14627; email: me@erincherry.net; Celine Latulipe, University of North Carolina at Charlotte, Department of Software & Information Systems, 9201 University City Blvd., Charlotte, NC 28223; email: clatulip@uncc.edu.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2014 ACM 1073-0516/2014/05-ART21 $15.00

DOI: http://dx.doi.org/10.1145/2617588

The Creativity Support Index



12 statements, two for each of six factors

15 paired comparisons: “When doing this task, it's most important that I'm able to…”

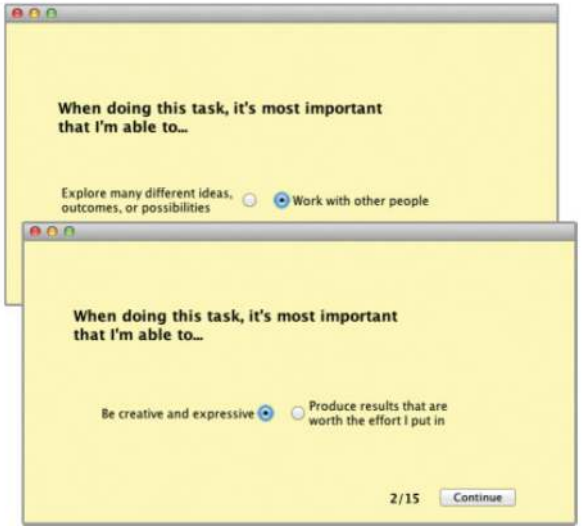
One score out of **100**

Cherry & Latulipe, TOCHI 2014, Fig. 1 and Sections 4.1 to 4.3

The paired comparisons

Table III. The Paired-Factor Comparison Test Has 15 Comparisons
For each pair, a user will choose a factor description in response to the following statement: "When doing this task, it's most important that I'm able to. . ."

1. Be creative and expressive
2. Become immersed in the activity
3. Enjoy using the system or tool
4. Explore many different ideas, outcomes, or possibilities
5. Produce results that are worth the effort I put in
6. Work with other people



Every pair of the six factors comes up once: **15** pages, one pair per page

A factor's count, from 0 to 5, is how many times it was picked.

Cherry & Latulipe, TOCHI 2014, Table III and Fig. 2

Scoring the CSI

$$CSI = [(Collaboration1 + Collaboration2) * CollaborationCount + (Enjoyment1 + Enjoyment2) * EnjoymentCount + (Exploration1 + Exploration2) * ExplorationCount + (Expressiveness1 + Expressiveness2) * ExpressivenessCount + (Immersion1 + Immersion2) * ImmersionCount + (ResultsWorthEffort1 + ResultsWorthEffort2) * ResultsWorthEffortCount] / 3.0$$

Each factor: its two statements added up, times how many times it was picked

Above **90**: an A. Below **50**: an F.

Cherry & Latulipe, TOCHI 2014, Fig. 3 and Section 4.3

The CSI for Google Docs

Table IV. CSI Results from a Collaborative Creative Writing Study Using Google Docs (N = 16)
The average CSI score for Google Docs in this study was 87.73 (SD = 11.30).

Scale	Avg. Factor Counts (SD)	Avg. Factor Score (SD)	Avg. Weighted Factor Score (SD)
Results Worth Effort	2.50 (1.41)	17.70 (2.30)	44.13 (27.38)
Exploration	2.56 (1.63)	16.00 (2.80)	41.00 (27.84)
Collaboration	3.63 (1.59)	18.13 (2.16)	67.63 (29.57)
Immersion	1.94 (1.34)	15.40 (4.94)	34.00 (26.00)
Expressiveness	3.19 (1.33)	17.63 (2.47)	55.44 (22.17)
Enjoyment	1.19 (0.91)	18.10 (2.19)	21.00 (16.69)

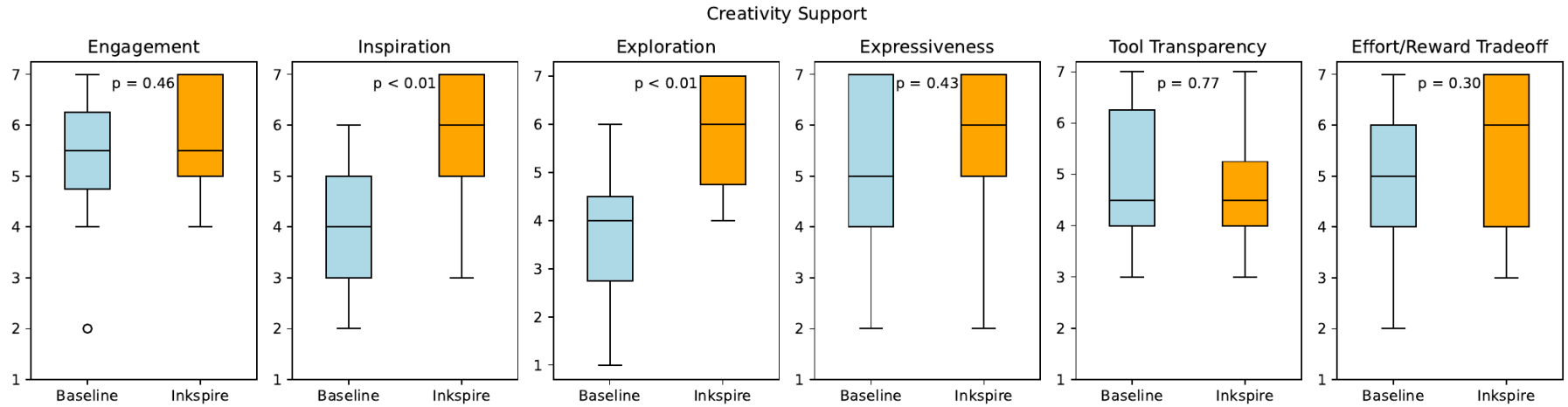
16 novices writing a short story together

Picked most: collaboration and expressiveness.
Picked least: enjoyment.

Change only one of tool, task or users at a time.

Cherry & Latulipe, TOCHI 2014, Table IV and Sections 4.4 and 4.5

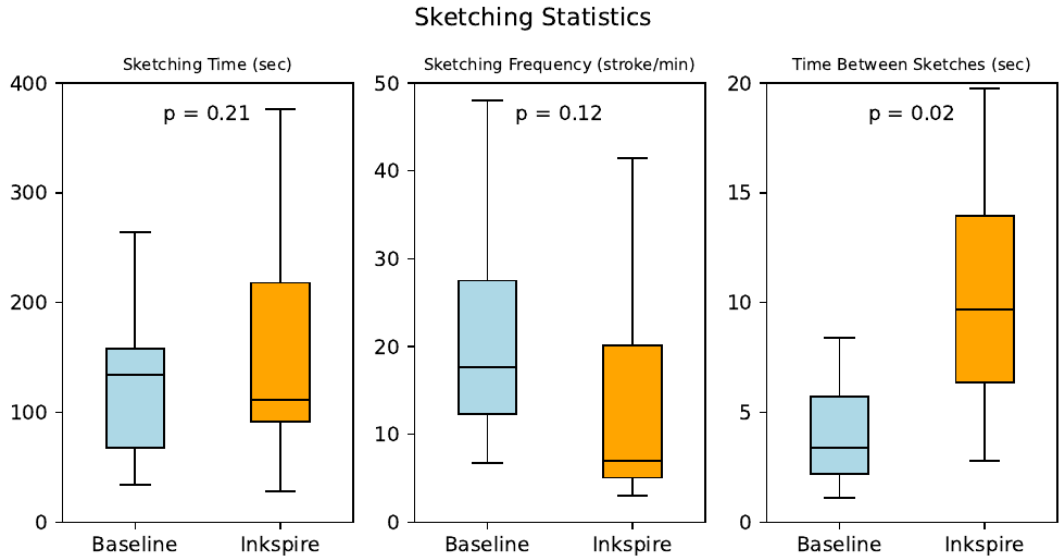
The CSI in a study of Inkspire



Inspiration and exploration rated higher ($p < 0.01$)
The other four factors not significantly different

Lin et al., CHI 2025, Fig. 5. Twelve participants, Inkspire against a ControlNet baseline, factors adapted from the CSI on a 7-point scale

Inkspire: measures from the logs



Prompts per participant: **8.5** with Inkspire, **11.9** with the baseline

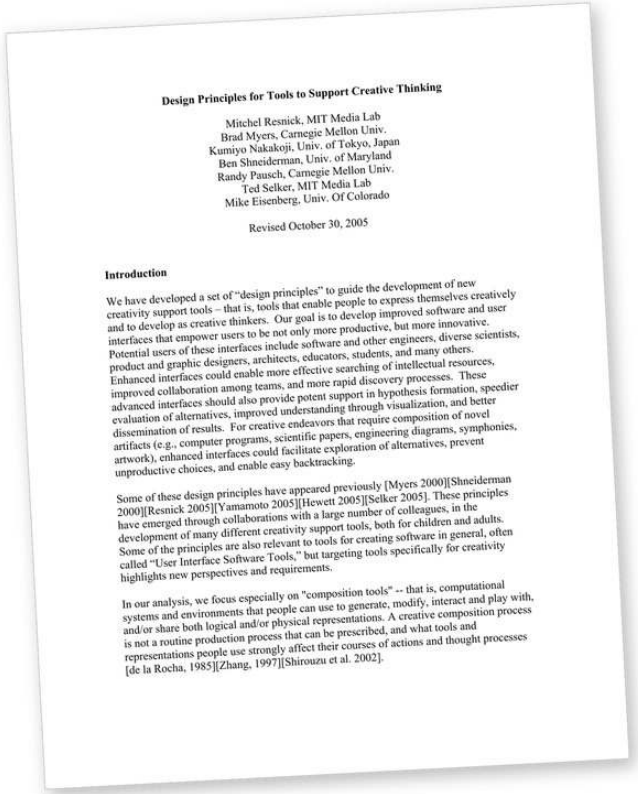
How alike one participant's prompts were (BERTScore): **0.51** with Inkspire, **0.76** with the baseline

Longer time between sketches with Inkspire (p = 0.02)

Lin et al., CHI 2025, Fig. 9 and Section 6

3. Design principles

Today's reading: Resnick et al., “Design Principles for Tools to Support Creative Thinking,” 2005



Design Principles for Tools to Support Creative Thinking

Mitchel Resnick, MIT Media Lab
Brad Myers, Carnegie Mellon Univ.
Kumiyo Nakakoji, Univ. of Tokyo, Japan
Ben Shneiderman, Univ. of Maryland
Randy Pausch, Carnegie Mellon Univ.
Ted Selker, MIT Media Lab
Mike Eisenberg, Univ. Of Colorado

Revised October 30, 2005

Introduction

We have developed a set of “design principles” to guide the development of new creativity support tools – that is, tools that enable people to express themselves creatively and to develop as creative thinkers. Our goal is to develop improved software and user interfaces that empower users to be not only more productive, but more innovative. Potential users of these interfaces include software and other engineers, diverse scientists, product and graphic designers, architects, educators, students, and many others. Enhanced interfaces could enable more effective searching of intellectual resources, improved collaboration among teams, and more rapid discovery processes. These advanced interfaces should also provide potent support in hypothesis formation, speedier evaluation of alternatives, improved understanding through visualization, and better dissemination of results. For creative endeavors that require composition of novel artifacts (e.g., computer programs, scientific papers, engineering diagrams, symphonies, artwork), enhanced interfaces could facilitate exploration of alternatives, prevent unproductive choices, and enable easy backtracking.

Some of these design principles have appeared previously [Myers 2000][Shneiderman 2000][Resnick 2005][Yamamoto 2005][Hewett 2005][Selker 2005]. These principles have emerged through collaborations with a large number of colleagues, in the development of many different creativity support tools, both for children and adults. Some of the principles are also relevant to tools for creating software in general, often called “User Interface Software Tools,” but targeting tools specifically for creativity highlights new perspectives and requirements.

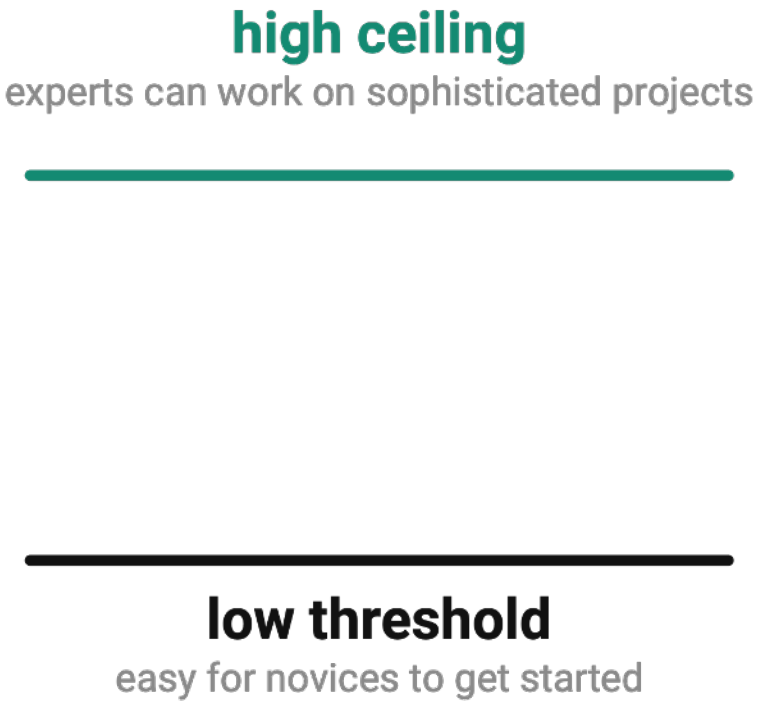
In our analysis, we focus especially on "composition tools" -- that is, computational systems and environments that people can use to generate, modify, interact and play with, and/or share both logical and/or physical representations. A creative composition process is not a routine production process that can be prescribed, and what tools and representations people use strongly affect their courses of actions and thought processes [de la Rocha, 1985][Zhang, 1997][Shirouzu et al. 2002].

Low threshold



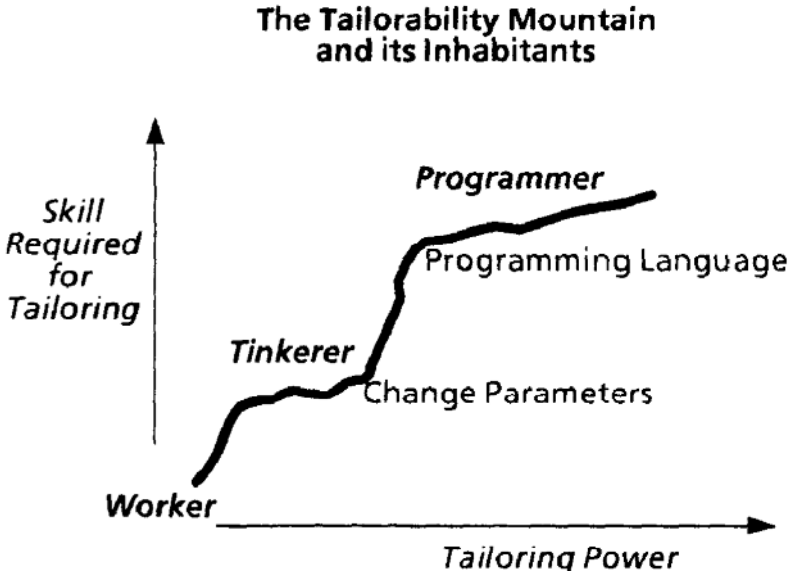
Resnick et al., 2005, Section 2

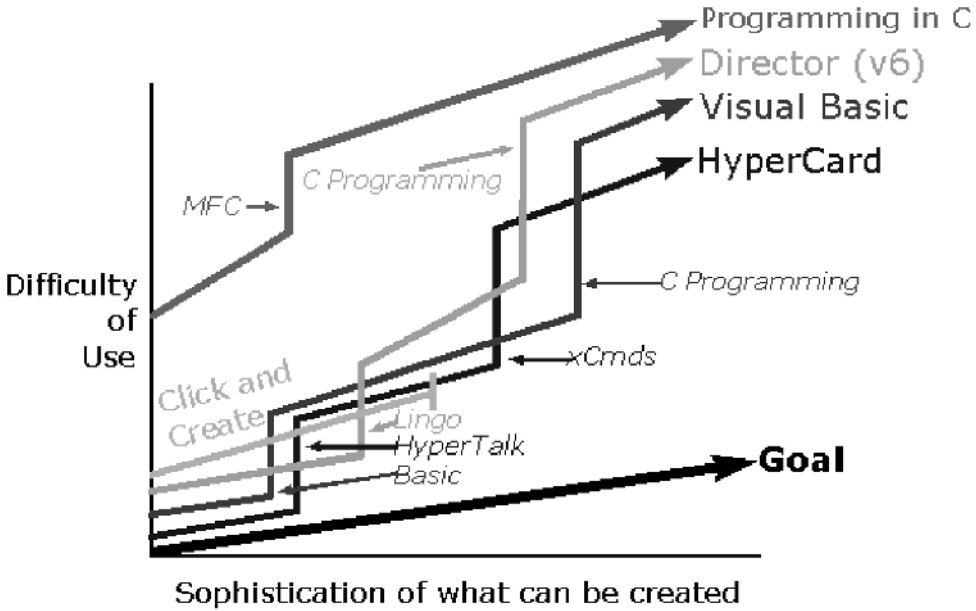
Low threshold, high ceiling



Resnick et al., 2005, Section 2

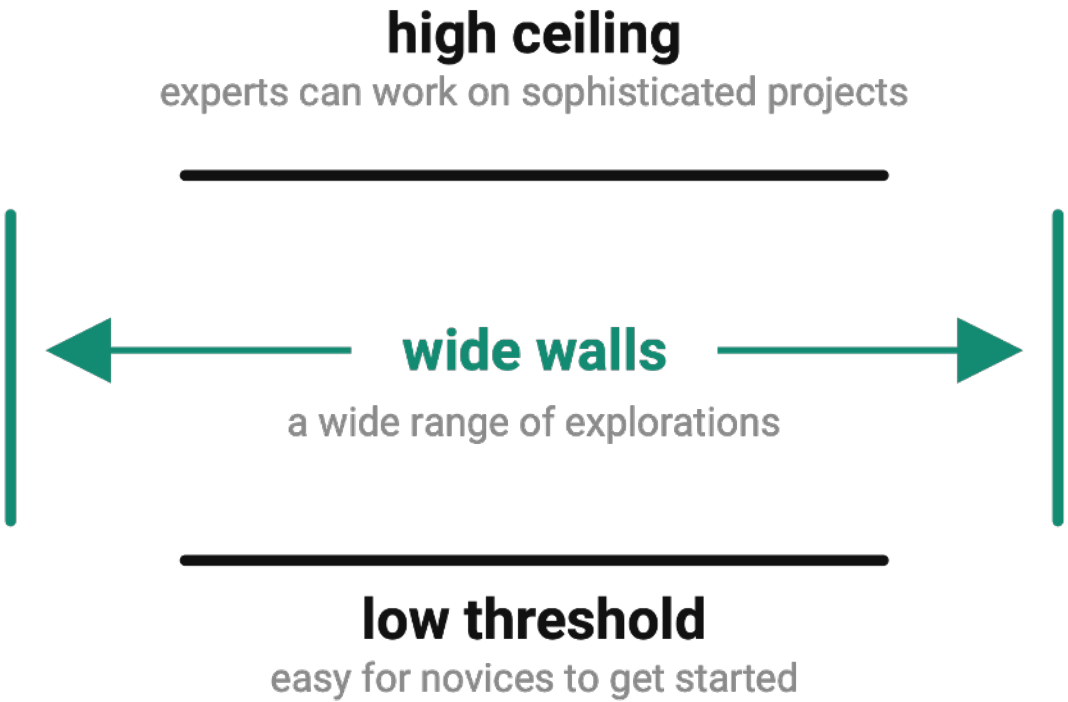
Low threshold, high ceiling





MacLean et al., CHI 1990, Fig. 1 (left); Myers, Hudson & Pausch, TOCHI 2000, Fig. 1 (right)

Low threshold, high ceiling, **wide walls**



Resnick et al., 2005, Section 2

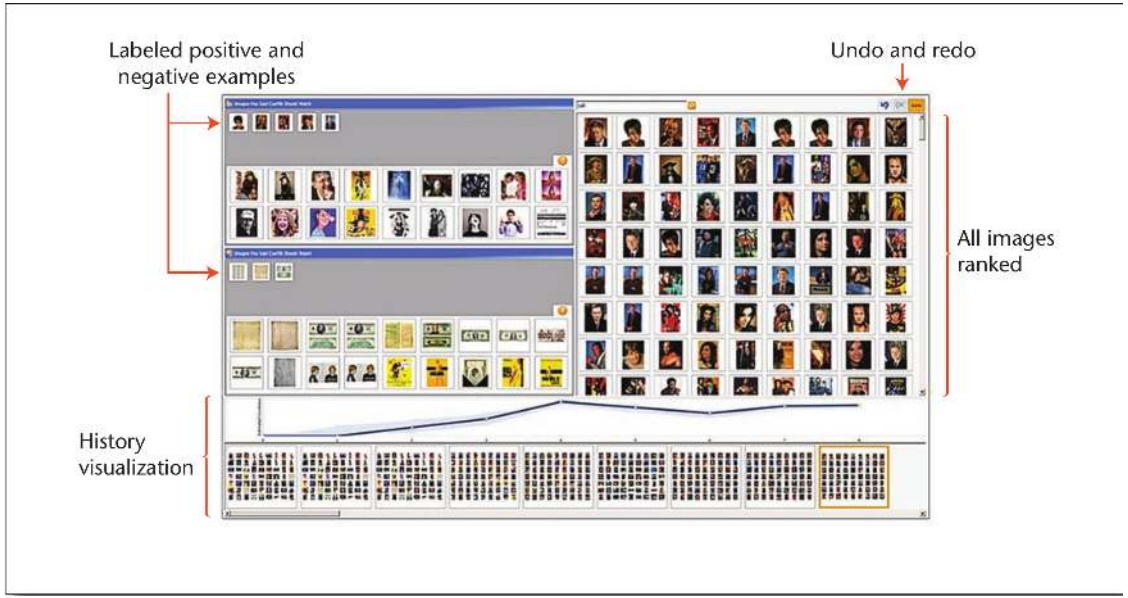
Low threshold, high ceiling, **wide walls**

“If the creations are all similar to one another, we feel that **something has gone wrong**.”

Resnick et al., 2005, Section 2

Support exploration

First, it must be very easy to try things out, and then backtrack when unsuccessful. This means that the tools must be trustworthy so that users are comfortable trying things. For example, a very good Undo capability is required in the tools.



CueFlik, from week 5: a history of earlier models, with undo and redo

Resnick et al., 2005, Section 1
Amershi et al., AI Magazine 2014, Fig. 6

Choose black boxes carefully

Similarly, for the language for programming the Cricket, the original design used the conventional Red, Green, Blue parameters to **setcolor.** This representation provides great power and flexibility, but kids found it difficult to map inputs from a single sensor into meaningful color values (for example, red for low values, blue/purple for high values). So a simpler **setcolor** command was added with just a single input that ranges from 0 to 100 (with 0 at the red end of the spectrum, 100 at the blue end), making it easy for kids to map sensor inputs to colors along the spectrum.



−1.5 −0.75 0 +0.75 +1.5

Our slider: one number, s

Hidden: the three predictions at every step

Resnick et al., 2005, Section 7; our slider from week 5

Evaluation of tools

12. Evaluation of Tools

One important issue with the design of creativity support tools is how they can be evaluated (see this report's section titled "Creativity Support Tool Evaluation Methods and Metrics"). How do you know if a tool is being helpful? Human-computer interaction professionals are used to measuring the effectiveness and efficiency of tools, but how do you measure if it supports creativity? As discussed above, tools that are *not* effective and efficient will probably hinder creativity, but it isn't clear that the reverse will hold. To try to measure creativity, the Silk system designers evaluated many different properties, including the number of different designs produced, the variability of the components used, the variety of questions about the designs from collaborators, etc. [Landay 1996], but these still do not really get at the *quality* of the solution. It is still an open question how to measure the extent to which a tool fosters creative thinking. While the rigor of controlled studies makes them the traditional method of scientific research, longitudinal studies with active users for weeks or months seem a valid method to gain deep insights about what is helpful (and why) to creative individuals

Resnick et al., 2005, Section 12

Discussion

1. Which CSI factors matter for your partner's task?
2. Who can start using your tool in the first minute?
3. How would you show it changed what designers made?

Led by this week's discussion leads, with questions from the reading posts

Evidence for the bake-off

Ratings: the three statements, and the CSI factors that matter for the task

Logs: prompts, undos, time between steps

Designs: what designers made, next to what they made without your tool

Technical: how fast it is and how the model behaves, like our slider table

Workshop: plan your bake-off test

1. Pick the real task your partner's designers will do
2. Choose your evidence: ratings, logs, designs or technical
3. Try the test with another team
4. Revise your test

Next week: the bake-off, with the final paper and videos due that day

Crit

1. Each team shows its tool for 3 minutes
2. Two teams and I answer these three questions
3. Write down three changes, and report back at the bake-off

Could I direct it? Could I see why it made what it made? Could I revise what it made?

Next: Week 10

Final Bake-Off

Bring your tool, running on your own laptop

Hand in: project video, process video, 4–6 page paper, public code repository