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

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.

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

Specifically, users provide ratings for six dimensions of creativity support: Enjoyment, Exploration, Expressiveness, Immersion, Results Worth Effort, and Collaboration. The CSI was inspired by the survey structure of the popular NASA Task Load Index (TLX) [Hart and Staveland 1988], which is a survey metric that assesses workload associated with completing a specific task. The TLX was used to help develop better tools and systems to decrease workload across a variety of dimensions. In contrast, the CSI measures how well a system or tool supports creative, open-ended activities, and it is grounded in literature about creativity support tools, creativity, play, and flow. Since the initial two CSI publications in 2009 (the beta version was presented as a CHI poster [Carroll and Latulipe 2009] and a revised version was presented at Creativity & Cognition [Carroll et al. 2009]), we have extensively developed and validated the CSI through a rigorous psychometric process. Over the course of the CSI's development, it has been used in studies with 13 different CSTs and over 120 participants.

This article is organized as follows: Section 2 introduces CSTs and the relevant literature on CST evaluation. Section 3 provides an overview of the NASA Task Load Index, which was the inspiration for our development of the Creativity Support Index. In Section 4, we present the final version of the Creativity Support Index. In doing so, we discuss how to administer and score the CSI and how to report CSI results, and we present a worked example of using the CSI in an experimental design. In Section 5, we give an overview of the development process of the CSI and present user studies that have been performed to validate the metric. Section 6 provides a detailed discussion of the Collaboration factor within the CSI and the final study that we performed to ensure that the CSI accurately reflects collaboration support, when applicable. Section 7 presents a discussion of scenarios in which researchers may want to modify the CSI, and we present guidelines for which modifications are acceptable and how such modifications should be handled. We conclude with a description of our future work directions and a summary of our contributions. The Appendix describes the rigorous psychometric development process that we used in developing the CSI; this process detail should be useful for other HCI researchers wishing to develop standardized survey metrics.

2. BACKGROUND

2.1. Creativity Support Tools

A creativity support tool is any tool that can be used by people in the open-ended creation of new artifacts. One could consider a set of paints, brushes, and a palette a creativity support tool for painting. One could consider a piano, score sheets, and a pencil the creativity support tool necessary for certain types of musical composition. And a typewriter might be the creativity support tool used by 20th-century writers. In the computing domain, CSTs are often software applications that are used to create digital artifacts or are used as part of the process of working toward the completion of an artifact. For example, a digital painter may use nothing but Adobe Illustrator to create an artwork. A designer may use a camera to photograph something and then edit, compose, and layer that photograph with other graphical imagery in Adobe Photoshop to create a final advertisement that is a physical artifact displayed on a billboard. CSTs span many different domains, and they support a variety of open-ended creative activities, including tools designed to support programming, information exploration, data analysis, and artistic creations in visual, performing, or musical arts.

Creativity support tools fall into a larger class of systems called creativity support environments (CSEs). While CSTs are traditionally desktop applications or a single piece of software, CSEs are more inclusive, ranging from environments that may require specialized hardware and instrumented spaces to environments that may exist

Table I. A Summary of Creativity Support Tools, Including Examples from Research and Industry

Category	Example
Visualization & Simulation	Tableau, D3, netLogo
Concept Mapping & Brainstorming	combinFormation, Omnigraffle, Whiteboards, Post-It Notes
Architecture & Design	AutoCAD, Rhino3D
Mathematics	SPSS, MatLab, WolframAlpha
Software Development Environments	Eclipse, Visual Studio
Video Editing	Final Cut Pro, iMovie
Drawing/Painting	Illustrator, InkScape, CorelDraw, Paper & Pen
Animation	Flash, Maya, SoftImage, Houdini
Music	GarageBand, Zya, Sequel, NodeBeat
Photography	Photoshop, Lightroom
Wikis, Blogs, & Online Presence	MediaWiki, WordPress, DreamWeaver
Writing & Presentation	Google Docs, MS Word, Prezi

purely within a collaborative space. Many new CSEs employ tangible devices that work in concert with digital software (such as Lego Mindstorms [Guilford 2001]). Our research is relevant to the larger class of CSEs, but it is focused primarily on CSTs. In Table I, we have compiled a summary of the types of creativity support tools, which builds on Shneiderman's [2007] summary.

It is also interesting to note that many software applications and digital systems that are thought of as creativity support tools are also simultaneously thought of as productivity support tools. This makes sense, as these tools often support a user across various iterative phases of a creative process: ideation, execution, and evaluation [Candy 2013]. In fact, an individual or team may use several of these tools simultaneously or in sequence during a creative work process. The categorization of productivity support tool or creativity support tool may also be dependent on the task. For example, a writer who uses Microsoft Word to write a short story is using MS Word as a creativity support tool, while an administrative clerk who uses Microsoft Word to mail merge form letters for company business is using MS Word as a productivity tool. It may be simple to evaluate how well Microsoft Word supports mail merge, as you could measure how long it takes the administrative assistant to create 100 form letters. Productivity tasks tend to be well defined and can be measured with straightforward quantitative metrics. However, the same metric could not be applied to measure how well the writer is supported by Microsoft Word in his or her task of short-story writing. This is the impetus for the development of the CSI.

2.2. Evaluating Creativity Support Tools

There are decades of research on defining and evaluating creativity [Albert and Runco 1999], but evaluating creativity support tools is a much newer field of study [Shneiderman et al. 2006; Shneiderman 2007]. According to Shneiderman, the main challenge in evaluating CSTs is that there are no obvious metrics to quantify, in contrast to productivity support tool evaluation, in which performance, time, and error rate are highly standardized measures.

Research on divergent thinking and creative ideation (or brainstorming) has also been influential to CST evaluation [Hocevar 1979; Hocevar and Bachelor 1989]. For example, Kerne et al. [2008] contributed an evaluation framework for studying information-based ideation tasks using combinFormation, a tool they developed for information discovery. These ideation tasks involve creative innovation, in which users' goals are to develop new ideas. The evaluation approach, called the Emergence Metric, involves judging the quality of synthesis in the generation of new ideas, and their novelty, and counting the number of ideas created.

A different approach is to study people while they are using a CST. For example, Kim and Maher [2005] developed a protocol analysis method for evaluating collaborative design and comparing graphical and tangible user interfaces. When evaluating creativity support tools, Hewett et al. [2005] noted that there is no one-size-fits-all approach, and because of this, researchers need to use a variety of metrics to show convergence. In developing the Creativity Support Index (CSI), we designed it to be flexible, so that it could be utilized to study a variety of CSTs and incorporated into many experimental designs. The CSI was designed with the intention of it being an additional evaluation metric that researchers could use in concert with other evaluation approaches.

3. RESEARCH INSPIRATION: NASA TASK LOAD INDEX

In developing the Creativity Support Index, we were inspired by the form factor of the NASA Task Load Index (TLX), which is a standardized survey used to quantify workload [Hart and Staveland 1988]. Similar to creativity, workload is a complex phenomenon that is understood intuitively but is not easily captured by any single metric. The TLX was originally developed for evaluating workload tasks in aircraft simulations and other similar human–machine equipment. Since its development over 20 years ago, the TLX has been reliably used (and even adapted) [Hart 2006] and has also been employed in a variety of domains, including the HCI community [Hewett et al. 2005; Latulipe et al. 2006]. Because of its widespread use, many researchers are already familiar with the TLX, which means that the TLX has meaning to researchers and the results can be reported without excessive narrative or explanation.

The NASA Task Load Index measures six factors of workload: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. There is one question per factor for a total of six questions. For example, the survey item for Frustration is, “How insecure, discouraged, irritated, stressed, and annoyed were you?” All questions are on a 21-point scale of “Very Low” to “Very High.” In addition to these six questions, there is a paired-factor comparison test, in which participants must compare each factor against every other factor for a total of 15 comparisons. In the paired comparisons, participants are asked to indicate which factor in a pair contributed the most to their workload. These comparisons are used to produce a weighted index score out of 100, where a higher score indicates a higher (more difficult) task load. The weightings from the paired factor comparisons make the overall index score more sensitive to the factors that are the most applicable in a given situation (i.e., some tasks may have high mental demand and low physical demand). By using weights relevant to the task, the survey generalizes across domains.

The results from the NASA TLX depend on the particular task being performed. Variations of the index score across different tools used to complete the task can be used to assess which tool or system allows the task to be completed with the least overall effort. It is also likely that the TLX score is somewhat dependent on the expertise of the person completing the survey: one would anticipate that a person who is an expert at a task will rate the mental demand on that task lower than a person who is a novice at that task. Thus, an individual TLX score is representative of three things: the task, the tool, and the user. For this reason, a single TLX score is not particularly meaningful. However, averaging TLX scores for a particular tool and task across a set of similar users (i.e., all experts vs. all novices) provides a reliable measure of task load for that task with that tool. When used in repeated measures studies where similar users perform the same task with different tools, a valuable comparative analysis emerges. In addition, a category-level comparison (i.e., examining physical demand between two systems with different input devices) can help to isolate what aspects contributed the most to the task load for different systems or tools.

Given the flexibility and power of the NASA Task Load Index, we were inspired to model the Creativity Support Index after it. It is our hope that as more researchers employ the CSI and report index scores, the metric will become as meaningful and useful to researchers as the NASA TLX.

4. THE CREATIVITY SUPPORT INDEX

The Creativity Support Index is a psychometric survey that we designed to assess the ability of a digital creativity support tool to support the creative process of its users. Its theoretical foundation is based on concepts from creativity and cognition support tools, which includes Boden's work on creative exploration and play [Boden 2004], formal theories of play [Read et al. 2002], Csiksentmihalyi's flow [Csikszentmihalyi 1997], and Shneiderman's design principles for creativity support tools [Shneiderman 2007].

In this section, we present our final version of the Creativity Support Index in order to make it available to other researchers. In doing so, we also discuss the administration of the CSI, provide a worked example from a creativity study, and provide a comparative study scenario with possible design implications, to help illustrate how the CSI results can be applied to the improvement of CSTs. We also present various scenarios for employing the CSI in different types of HCI study designs.

4.1. The CSI: Final Version

The Creativity Support Index is very similar in structure to the NASA Task Load Index in that it consists of a rating scale section and a paired-factor comparison section. There are six factors: Collaboration, Enjoyment, Exploration, Expressiveness, Immersion, and Results Worth Effort. For each factor, there are two agreement statements, which differs from the TLX, where there is one agreement statement per factor. Having two statements per factor improves the statistical power of this survey, as it allows researchers to look at the reliability for each factor by examining the similarity of the scores across the two different statements. The agreement statements are shown in Table II. Each statement should be rated by participants on a scale of "Highly Disagree" (0) to "Highly Agree" (10). A screenshot of these agreement statements from the CSI's electronic version is shown in Figure 1.

Also similar to the NASA TLX, the paired-factor comparison section consists of each factor paired against every other factor for a total of 15 comparisons. In these comparisons, a factor description is selected in response to the statement: "When doing this task, it's most important that I'm able to..."" These six factor descriptions are available in Table III. Figure 2 shows two screenshots of the paired-factor comparisons from the electronic version of the CSI.

4.2. Administration

When administering the CSI, participants should complete the agreement statement section for each creativity support tool that they use. The paired-factor comparison should be administered for every task that they complete. Therefore, if a study involves one task with two different tools, participants should complete the agreement statement section twice (i.e., once after each CST) and complete the paired-factor comparison at the very end. Of course, this setup assumes that each user is performing the same task with each tool. If the user is not performing the same task with each tool being studied, then the researcher should administer the paired-factor comparison section after each tool. In Section 4.6, we discuss other scenarios in which researchers may employ the CSI. Participants can easily complete the CSI within 5 minutes; it does not add significant time to a research study.

Table II. These Are the 12 Agreement Statements on the CSI

Each agreement statement is answered on a scale of “Highly Disagree” (1) to “Highly Agree” (10). In deployment, the factor names are not shown, and the participant does not see the statements grouped by factor.

Collaboration

1. The system or tool allowed other people to work with me easily.
2. It was really easy to share ideas and designs with other people inside this system or tool.

Enjoyment

1. I would be happy to use this system or tool on a regular basis.
2. I enjoyed using the system or tool.

Exploration

1. It was easy for me to explore many different ideas, options, designs, or outcomes, using this system or tool.
2. The system or tool was helpful in allowing me to track different ideas, outcomes, or possibilities.

Expressiveness

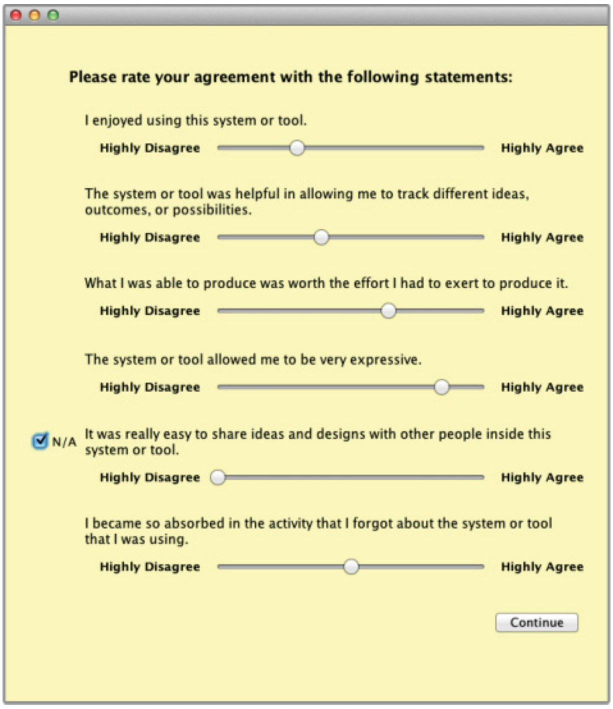
1. I was able to be very creative while doing the activity inside this system or tool.
2. The system or tool allowed me to be very expressive.

Immersion

1. My attention was fully tuned to the activity, and I forgot about the system or tool that I was using.
2. I became so absorbed in the activity that I forgot about the system or tool that I was using.

Results Worth Effort

1. I was satisfied with what I got out of the system or tool.
2. What I was able to produce was worth the effort I had to exert to produce it.



Please rate your agreement with the following statements:

I enjoyed using this system or tool.

Highly Disagree — Highly Agree

The system or tool was helpful in allowing me to track different ideas, outcomes, or possibilities.

Highly Disagree — Highly Agree

What I was able to produce was worth the effort I had to exert to produce it.

Highly Disagree — Highly Agree

The system or tool allowed me to be very expressive.

Highly Disagree — Highly Agree

☑ N/A It was really easy to share ideas and designs with other people inside this system or tool.

Highly Disagree — Highly Agree

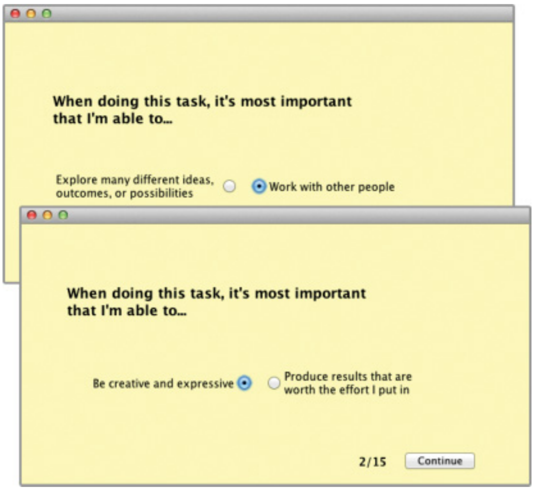
I became so absorbed in the activity that I forgot about the system or tool that I was using.

Highly Disagree — Highly Agree

Fig. 1. Screenshot of six of the 12 agreement statements in the CSI’s user interface.

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



When doing this task, it's most important that I'm able to...

Explore many different ideas, outcomes, or possibilities ○ ● Work with other people

When doing this task, it's most important that I'm able to...

Be creative and expressive ● ○ Produce results that are worth the effort I put in

2/15

Fig. 2. Screenshot of two of the 15 paired-factor comparisons in the CSI's user interface.

It is entirely possible to administer the CSI on paper; however, we recommend that researchers administer the CSI using the application that we developed.[1] The application will score each test automatically with results saved in a comma-separated file (csv), labeled with participant ID and condition number (i.e., in the case of repeated measures). In the CSI application, there is an initial page that allows the researcher to configure the survey administration by entering a participant ID and specifying how many times the CSI will be completed by the participant (depending on the number of conditions being tested or tools being used). The participant is then presented with two pages of agreement statements with six items per page, as seen in Figure 1. There are also 15 pages of paired-factor comparisons with one pair per page (Figure 2). We made the design decision to have one factor per page to force participants to consider each pairing independently and to prevent participants from lining up their answers to the paired comparisons. In both the paper and digital versions of the NASA TLX, the 15 paired comparisons were routinely presented together on one page in a list, which may have led to participants clicking down the list and not really considering each pair independently.

4.3. Scoring the CSI

The CSI application scores the surveys automatically, generating a single CSI score out of 100 for the tool being used, with a higher score indicating better creativity

[1] The CSI application is available for download at http://www.erincherry.net/csi.

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

Fig. 3. Equation for manually scoring the CSI.

support. This scoring system has a nice mapping to educational grading systems, and researchers can use grades as a shorthand for interpretation: a score above 90 is an "A," which indicates excellent support for creative work. A score below 50 is an "F," which indicates that the tool does not support creative work very well. The CSI also generates individual factor scores that can help a researcher understand how a tool supports various aspects of creative work. The scoring equation is shown in Figure 3 for researchers who deploy the CSI on paper and need to score the responses manually. To summarize the equation: the CSI is scored by first summing the agreement statements for each factor to get a factor subtotal. Each factor subtotal is then multiplied by its factor comparison count (i.e., the number of times it was chosen in the factor comparisons). Finally, these are summed and divided by three for an index score out of 100.

4.4. CSI Interpretation

Probably one of the most critical questions to ask about this metric is, "What does a CSI score for a given system mean?" The CSI score is a reflection of how well that tool supports creativity for the particular task or activity the user was engaged in and that is likely dependent on both individual preferences and the individual's level of expertise with the tool. As with most metrics, an individual score will reflect individual differences, and thus the most useful information (unless you are interested in scoring the individual) is achieved through aggregation. This is true with the NASA TLX scores, which tend to reflect the task as well as the average level of expertise of the individuals completing the survey. The CSI reflects tool, task, and expertise level of the user. We would anticipate that the CSI score for the same task and tool would be different if the survey was completed by a group of novice participants rather than a group of expert participants. We would expect a CSI score for the same tool and a group of experts at that tool to differ if half the experts were given one creative activity and half the experts were given a different creative activity. And finally, if we had a group of experts all doing the same task but using two different tools, the CSI scores again should be different. The key is to ensure that only one of the factors (tool, task, and expertise of participant group) varies at a time; otherwise, comparison of the CSI score between two different treatments will be difficult to decipher. It is also important to explicitly report the conditions (task performed, level of expertise of users) under which a CSI score was obtained.

A CSI score for a CST is not necessarily representative of the whole CST; rather, it reflects that CST being used for a particular activity, by a particular type of user. We are unlikely to report an overall CSI score for a complex product such as Adobe Illustrator, because that would be meaningless. What would be meaningful is a CSI score for Adobe Illustrator for the activity of drawing comics by novice users, or a CSI score for Adobe Illustrator for the activity of digital sketching by expert users. There is a further distinction, which is whether the level of expertise refers to expertise in the

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)

domain or to expertise with the tool, and that could also impact the results. Again, this means that it is important for researchers to explicitly identify which kind of expertise the participants have. We believe that the CSI will be more affected by the expertise participants have with the tool than by their domain expertise; however, that is left for further research.

4.5. CSI Example

Here we provide a worked example of how the CSI can be deployed, interpreted, and reported. This example is from a collaborative creative writing study conducted by the authors, in which novice creative writing participants used Google Docs to collaboratively write a short story in response to two photographic prompts. The study itself is discussed further in Section 6.

Novice participants generated an average CSI score of 87.73 (SD = 11.30) for Google Docs in a collaborative creative writing study. Table IV shows the average factor counts, average factor score, and average weighted factor score for each of the six factors on the CSI. The average counts are the number of times that participants chose that particular factor as important to the task (see Table III for the factor comparisons). Note that the average factor counts also tell us something about which factors are most important for this particular creative activity (regardless of how well the tool supported those factors). The highest possible count for any particular factor is 5, indicating that participants chose it as more important than every other factor. In particular, immersion and enjoyment do not seem as relevant or important to users engaged in collaborative writing, while collaboration and expressiveness are very important (Table IV).

The factor score represents the sum of both agreement statement responses for a factor, each of which is on a scale between 0 and 10, with a higher number indicating that the tool better supports that factor. Therefore, the maximum factor score is 20. In Table IV, we see that all factors received relatively high ratings, with Collaboration, Enjoyment, and Expressiveness being rated the highest. However, since Enjoyment was not rated particularly important to this task, we see that Enjoyment receives a lower-weighted factor score. Weighted factor scores are calculated by multiplying a participant's factor agreement scale score by the factor count, in order to make the weighted factor score more sensitive to the factors that are the most important to the given task.

This particular study was not a comparative study where participants also used another tool for collaborative writing. For the sake of illustration of comparative analysis, let us imagine that participants in the Google Docs study had also been asked to collaboratively write a short story using CollabText, a (fabricated for illustration) text collaboration tool. The results from this scenario are shown in Table V. In such a within-subjects study, the participants would do the factor comparison part of the CSI

Table V. **Imagined** CSI Results Using CollabText for a Collaborative Creative Writing Study
In this imagined scenario, CollabText would receive an overall CSI score of 81.7.

Scale	Avg. Factor Counts (SD)	Avg. Factor Score (SD)	Avg. Weighted Factor Score (SD)
Results Worth Effort	2.50 (1.41)	19.34 (2.10)	48.35 (22.73)
Exploration	2.56 (1.63)	14.04 (2.25)	35.94 (24.02)
Collaboration	3.63 (1.59)	19.45 (0.87)	70.60 (25.31)
Immersion	1.94 (1.34)	16.20 (1.35)	31.43 (18.23)
Expressiveness	3.19 (1.33)	13.02 (1.89)	41.53 (21.31)
Enjoyment	1.19 (0.91)	14.50 (3.54)	17.25 (20.13)

based on the task of collaborative writing, and so the factor counts would be the same as in Table IV. However, the participants would give statement agreement ratings for CollabText separately from the statement agreement ratings given for Google Docs, and this would account for any differences in scoring.

If the factor scores in Table V were real, one would note that the overall CSI scores are similar, with Google Docs having a slightly higher overall score (87.73 vs. 81.7). Both of these can be considered good creativity support scores (“B” grades), but not excellent. This means that both of these tools would provide reasonable creativity support to users engaged in collaborative creative writing, but each has room for improvement.

In order to determine where best to target design efforts to increase creativity support, we could further break down the analysis and look at the individual factors and how they were rated across the two applications.

Results Worth Effort. The average count for the results worth effort factor is 2.5, suggesting it is of moderate importance to users engaged in collaborative writing. The Google Docs score for this factor is 17.7; the CollabText factor has a score of 19.34. This suggests that the amount of effort required for the same amount of work is higher with Google Docs, and that there may be some interactions that are tedious, or take too many steps. In such a case, software log analysis may help to reveal repeated series of actions where an interaction redesign could reduce the effort required.

Exploration. The average count for the exploration factor is 2.56, suggesting it is also of moderate importance in collaborative writing. The Google Docs score for exploration is 16, while the CollabText score is 14.04. This suggests that it is easier to consider different possibilities or versions or to try out new ideas with Google Docs than with CollabText. The CollabText designers might want to consider investigating their support for document versioning, branching, and duplication.

Collaboration. As expected, the average count for collaboration is 3.63, which indicates support for collaboration is important to users engaged in collaborative writing. The Google Docs score for collaboration is 18.13, while the score for CollabText was 19.45. Here we also note a very low standard deviation for the CollabText collaboration factor score, indicating very low variability across participants on this scale. This suggests that CollabText does a very good job of supporting collaboration. In this case, Google Docs designers may want to consider evaluating aspects of collaboration support in their editor, such as the support for in-document chat, collision detection, visual representations of how recently parts of text were edited by others, and so forth.

Immersion. The average factor count for immersion was 1.94, which suggests it is less important to users engaged in collaborative writing, and the statement agreement scores are not particularly different across the two applications. However, there is a very high standard deviation on the Immersion score for Google Docs, indicating that there was high variance across users on this particular scale. This means that researchers should be careful interpreting and acting on these results, as they are not as strong as the other rating results. If one of the applications had a particularly low score on the immersion factor, it would indicate that designers should look to reduce notifications and interruptions that may be disrupting the workflow.

Expressiveness. The average factor count for expressiveness is 3.19, which indicates that this is important to users engaged in collaborative writing. The Google Docs score for expressiveness is 17.63, while the CollabText score is only 13.02. This indicates that users are frustrated by support for expressiveness in CollabText. Perhaps the formatting in CollabText is too limited or the ability to write different versions of a story too constrained, or perhaps there is no thesaurus or dictionary support to help users search for words that would better express their thoughts.

Enjoyment. The average factor count for enjoyment is low, indicating that enjoyment is not particularly important to users when engaged in collaborative writing. The Google Docs score for enjoyment is higher than the enjoyment score for CollabText, so if the designers consider this important, they may want to investigate ways to enhance the visual design of CollabText and also attempt to make the interaction more fun, perhaps considering the implementation of game-like statistics that show how many words have been written in real time.

This worked example illustrates how the overall score or grade of a CST is useful for comparison purposes, but that it is the individual factor scores that can help researchers or designers know where to focus their efforts if they wish to improve the creativity support of a particular tool.

4.6. CSI Usage Scenarios

The CSI can be used in a variety of study designs, and here we provide some scenarios to illustrate these possibilities.

Tool Comparison, Same Task, Repeated Measures. Researchers are interested in comparing the creativity support of two similar tools, so they design a study in which participants complete the same creative task in both tools. In this case, participants will complete the CSI's agreement statements after completing the task in each tool. Since the task was the same for both CSTs, the participant will complete the paired-comparison section only once, at the end of the study.

Tool Comparison, Same Task, Between Groups. This is similar to the previous case, except that participants use only one tool. This may be appropriate if significant learning transfer is expected from one tool to the next. In this case, the participant will complete the entire CSI after using the assigned tool. The factor counts should be reported for each group, as they may differ.

Multiple Tasks in the Same Tool. A creativity support tool is being studied by researchers who are interested in understanding a tool's ability to support various creative activities. In this case, the researchers will compute CSI scores for each creative task being studied while participants use the CST. After each task, a participant will complete both the agreement statement section and the paired-comparison section. Participants must complete the paired-comparison section multiple times because the researchers are studying different tasks within a creativity support tool, and the paired-factor comparison test is focused on the task. The CSI data are analyzed by looking at the difference in CSI scores for each task. The researchers are also interested in comparing the results of the paired-factor comparison test between the different creative tasks, in order to understand whether certain factors were more important in one creative task versus another.

Longitudinal Study of a CST. Researchers are studying a CST through a longitudinal study in order to understand whether creativity support increases as participants develop more expertise with the tool. In this study, participants use the same CST and complete the Creativity Support Index at the end of each session. In this scenario, the researchers are interested in changes of the CSI scores over time, and they may also find interesting changes in particular factors over time. For example, as a participant becomes more adept at using a complex CST, their enjoyment ratings may increase.

Tool Without Collaboration. A researcher is studying a CST that does not support collaboration, and subsequently, there are no collaborative tasks. The researcher's first plan is to remove all the collaboration statements from the CSI. However, the researcher decides to administer the complete CSI, allowing participants to mark the collaboration statements as nonapplicable. This benefits the researcher for several reasons. First, this will allow the researcher to report an actual CSI score from a standardized survey metric, rather than having to adapt the metric and justify the adaptation. Second, the paired-factor comparison section will still ask participants whether collaboration was important to them for the creative task; therefore, it will inform the researcher in cases where collaboration is desired by participants.

No Comparisons. In evaluating a creativity support tool, a researcher uses the CSI as an additional research metric but is not interested in comparing CSI scores to another tool or to another creative task. It is simply administered as a postexperiment survey. In this case, the researcher calculates the CSI and reports it in his or her paper as a comparison metric for other researchers to use that are studying similar CSTs.

Individual Rating Scales. While the CSI provides an overall index of creativity support for a particular task and tool, researchers comparing two different CSTs may be most interested in differences between particular factor scales, such as Immersion and Results Worth Effort. In their user study, they administer the complete CSI, but in their analysis, they decide to also investigate statistical differences on the statements that ask about Immersion and Results Worth Effort.

5. CSI VALIDATION

We developed the CSI following a rigorous psychometric process similar to that used in the development of the NASA TLX. This involved creating a beta version based on creativity theory, running user studies with the beta version, improving the factor names through a Mechanical Turk study of creativity vocabulary, and then lengthening the CSI and performing a factor analysis to determine optimal statement wording. This iterative development is described in detail in the Appendix. Our previous publications describe the early studies [Carroll and Latulipe 2009; Carroll et al. 2009]; here we present the studies performed more recently in order to validate the metric.

5.1. Test-Retest Reliability

We conducted a study to test the stability and reliability of the CSI over time, using the two-item per factor version presented in Table IX of the Appendix. While test-retest studies are very common in psychometrics, these studies are more challenging for a metric like the CSI, since it must be administered after completing a creative activity with a CST. Given that people become more familiar with a tool over time, we expected

Table VI. Average, Standard Deviations, and Test-Retest Reliability from the Overall CSI Score and Each of the Six Scales from Session 1 and Session 2
Note that for the averages, the CSI is out of 100, and each individual scale is out of 20.

	Session 1	Session 2	Reliability
Results Worth Effort	14.18 (5.12)	14.36 (3.70)	0.695
Exploration	10.36 (3.98)	11.45 (4.23)	0.643
Collaboration	8.55 (4.46)	9.91 (4.30)	0.238
Enjoyment	13.45 (6.06)	13.27 (4.54)	0.806
Expressiveness	13.64 (5.44)	13.18 (6.19)	0.494
Immersion	11.27 (6.99)	11.54 (5.37)	0.382
Overall CSI Score	64.46 (24.29)	64.12 (21.44)	0.636

Table VII. Descriptive statistics and Cronbach's alphas for All Six CSI Scales for Study 1, Study 2, and Study 3 in Section 5.2
Also included are the average and standard deviations for all three overall CSI scores.

	Study 1		Study 2		Study 3	
	Avg (SD)	Alpha	Avg (SD)	Alpha	Avg (SD)	Alpha
Results Worth Effort	17.20 (4.21)	0.972	14.45 (4.61)	0.832	15.95 (3.24)	0.865
Exploration	15.80 (3.27)	0.467	12.27 (4.08)	0.519	15.40 (3.75)	0.786
Enjoyment	17.20 (4.15)	0.953	12.27 (3.85)	0.847	15.10 (4.32)	0.791
Expressiveness	17.40 (3.71)	0.870	14.82 (3.74)	0.839	15.90 (3.63)	0.864
Immersion	15.80 (5.36)	0.920	10.73 (4.52)	0.568	14.90 (3.86)	0.750
Collaboration	16.40 (4.16)	0.971	9.36 (3.08)	0.281	9.85 (1.98)	0.802
Overall CSI Score	84.20 (18.84)	*N/A*	64.79 (17.06)	*N/A*	76.52 (16.25)	*N/A*

to see slight changes in scores over time, unless using experts as participants or using very simplistic creativity support tools.

We recruited 12 participants with an interest in sketching to participate in this test-retest study. In recruitment, all participants were aware that the study involved returning 3 weeks later to repeat the same experiment. We used a very basic web-based drawing application in this study, called Odosketch. Since this tool has a small set of paintbrushes and colors to choose from, we expected that learning on this tool would be very minimal, which would help reduce variations in scores across sessions due to learning effects.

In the first session, participants were given a short demo of Odosketch and were asked to spend 30 minutes sketching anything that they liked. After sketching, each participant completed an electronic version of the CSI from Table IX. All of our participants came back 3 weeks later to complete the same study again. In this second session, the only procedural change was that a demo of Odosketch was not provided. We paid participants $5 for the first session and $10 for the second session. We allowed them to choose what they wanted to sketch, as we did not want to impair their ability to be creative by assigning them something to draw that they were not interested in.

Overall, our test-retest reliabilities were very good (Table VI). The reliability of the overall CSI score from Session 1 to Session 2 was 0.65. We also calculated the reliability of the individual scales by summing the two items on each scale and finding the reliability of the summed scores from Session 1 to Session 2. We found satisfactory reliability for all but one of our scales, which was Collaboration.

5.2. Other Studies

We also conducted several studies that incorporated the two-items-per-scale version of the CSI from Table IX. In Table VII, we have summarized the results for each study.

5.2.1. Study 1: Adobe Photoshop. We visited an adult education class on Adobe Photoshop that was taught at The Light Factory (a photography museum in Charlotte, NC) and had students fill out the CSI in response to the activity that they were working on for their course. There were five female participants between the ages of 25 and 55. The average CSI score for Adobe Photoshop in this study, where novice users were using the tool for general photographic postprocessing, was 84.20 (SD = 18.84).

5.2.2. Study 2: AutoDesk Sketchbook Express. We used the CSI as an additional measure in a study that investigated the temporal, creative work process, using self-report and physiological measures [Carroll and Latulipe 2012]. In this study (N = 11), the task was open-ended sketching with AutoDesk Sketchbook Express while wearing electrodermal activity (EDA) sensors and an electroencephalography (EEG) headset. The average CSI score for open-ended sketching with SketchBook Express was 64.79 (SD = 17.06) for the novice users in this study.

5.2.3. Study 3: Bimanual Color Exploration Plugin (BiCEP). The CSI was also used in a study to evaluate the Bimanual Color Exploration Plugin (BiCEP), which is a color chooser designed for Mac OS X that allows users to explore the color space with two fingers using a touchpad [Gonzalez and Latulipe 2011]. There were 16 participants in this study, and they completed the CSI after using BiCEP for an open-ended coloring activity. The average CSI score for BiCEP for a coloring activity was 76.52 (SD = 16.25).

5.3. Challenges with Collaboration

The poor test-retest reliability for the Collaboration scale ($r = 0.238$) indicated that measurement error occurred for this scale. The most likely reason for this result is that the test-retest study did not involve a collaborative activity. Since the assigned task was not collaborative and the CST did not explicitly support collaboration, several participants expressed concern about how to respond to the statements about collaboration. Some participants wanted to mark collaboration as “Not Applicable,” but this was not an option. Instead, many participants chose to leave the slider in the middle, probably assuming that the middle value served as a neutral point, which is not the case on this continuous rating scale. Given this ambiguity, it is likely that after 3 weeks went by, some participants answered this question differently than in their first session.

After this study, we were particularly concerned with the Collaboration scale and acknowledged that more work was needed in this area. Specifically, we decided to revisit the items in the Collaboration scale, allow for “N/A” responses in the CSI’s user interface, and test the CSI with a collaborative tool.

6. THE COLLABORATION FACTOR

The development process described in the Appendix and in the previous section allowed us to create an almost final version of the CSI, but the collaboration factor clearly needed deeper consideration and improvements. Although the CSI was tested across multiple tasks and tools, we had yet to test the CSI in a study that actually involved a collaborative activity and a tool that explicitly supported collaboration. This is partly due to the fact that CSTs, which support collaboration, are much less prevalent than single-user desktop tools. Since the CSI had yet to be tested on a collaborative activity and a CST explicitly supporting collaboration, it was likely that the Collaboration items were not truly representative of the best questions for the Collaboration scale. Therefore, in an effort to improve the Collaboration scale of the CSI, we conducted a final user study with the goal of finding the best Collaboration items by testing the CSI on a collaborative activity and tool. This section describes that study and the final changes to the CSI.

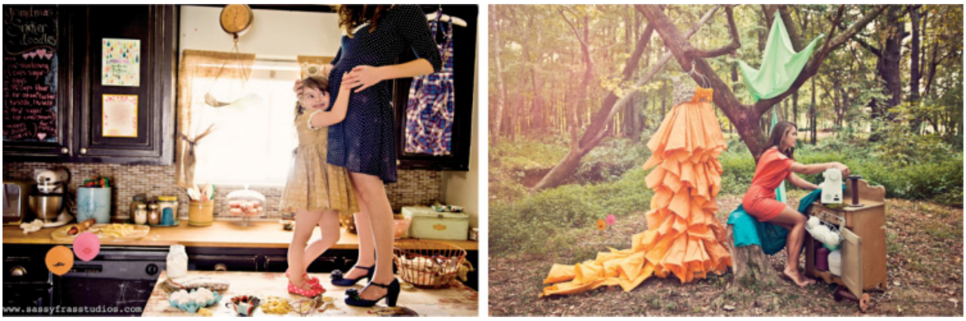


Fig. 4. Photographs used for the collaborative, creative writing task. Participants were instructed to write a story about how these photographs were connected using Google Docs as the writing tool. (Photos by Christen Lesley Lucas[2])

6.1. Collaboration Study

6.1.1. Methodology. We conducted a study that involved collaborative, creative writing using Google Docs. There were 16 participants in this study, who were recruited through Amazon's Mechanical Turk (MTurk). Each participant collaboratively wrote a creative story with a "confederate participant," rather than another study participant. In other words, our MTurk participants were led to believe that they were writing a creative story with another Turker. However, as a control variable, we opted to pair our participants with a confederate, rather than pair two study participants together. The confederate in this study was a student pursuing a Master of Arts in creative writing.

When people visited our study posting on MTurk, they were told that they would be collaboratively writing a creative story on Google Docs and that they would complete a short exit survey (e.g., the CSI) when they were finished writing. They were also told that it would take no more than 5 minutes to be matched up with a partner from MTurk. After accepting the task, the participant was given a link to our study's Google Doc, where the confederate participant was always waiting. We used a confederate participant in order to ensure that each participant had a good writing partner. However, the participants were under the impression that they would be collaboratively writing with another study participant. This deception was important, as we did not want to impose additional social pressure on the participants by telling them that they were writing with a member of our research team.

Inside the Google Doc, participants were instructed to collaboratively write for 30 minutes in response to a creative writing prompt. The writing prompt included two different photographs[2] that were unique and did not seem to belong together (Figure 4). They were instructed to "Tell the story of the connection between these photos. Choose a perspective to act as a voice, such as an external narrator or one of the characters in the photos." After writing, participants were instructed to take an exit survey, so a link to a web-based version of the CSI was also included inside this document.

The CSI that participants completed in this study was modified from the two-item-per-scale version in Table II. For collaboration, participants were given a total of four agreement statements. These additional collaboration items came from the extended version of the CSI described in Appendix A.3, when we did an item-level analysis for each of the six factors. Specifically, from that item-level analysis, we took the four collaboration items that performed the best (i.e., highest corrected item-total

[2] Photographer: Christen Lesley Lucas, http://www.sassyfrassstudios.com.

correlation). This allowed us to essentially perform a second, more accurate, item-level analysis on the Collaboration scale.

6.1.2. Analysis and Results. In order to finalize the Collaboration scale, we performed an item-level analysis on the four collaboration items used in this study. Similar to the item-level analysis in Appendix A.3, we first dropped the question with the lowest corrected item-total correlation and then we dropped the other question by looking at item content. Since only one of the questions was a negative question (i.e., reversed rating scale), we decided to drop that question. In this study, our Collaboration scale had a Cronbach's alpha of 0.914, and the selected questions were:

(1) It was really easy to share ideas and designs with other people inside this tool.
(2) The system or tool allowed other people to work with me easily.

6.2. Interface Adjustment for Collaboration Statement

Throughout our user studies with the CSI, participants were continuously confused by how to answer items about collaboration, when either the task was not collaborative or the tool did not explicitly support collaboration. When we began development of the CSI, we expected participants to mark "Highly Disagree" for statements in which collaboration was not involved, which would automatically score Collaboration statements as a 0 on a scale of 0-10. However, participants consistently told us that they wanted to mark this item as nonapplicable. Therefore, in the final version of the CSI, we added a check box beside the collaboration items, which allows participants to mark these items as nonapplicable (Figure 1). When the "N/A" check mark is ticked, the slider automatically repositions itself at "Highly Disagree," thus coding Collaboration statements as 0.

7. DISCUSSION

Sometimes researchers may want to modify the Creativity Support Index. However, for the CSI to be a standardized, psychometric tool, it went through a rigorous development process. If a researcher modifies the CSI, his or her results cannot easily be used in comparison with CSI results reported by others. Before modifying the CSI, we recommend that researchers reflect on these modification scenarios.

Adding CST's Name. The CSI was designed to be generic and not specific to one particular system or tool, as is evident in the wording of the CSI statements. For example, one of the Enjoyment statements is, "I would be happy to use this *system or tool* on a regular basis." Some researchers may prefer to actually change "system or tool" to the name of the CST they are studying, and this modification is perfectly acceptable.

Removing Collaboration. Many researchers will be studying CSTs that do not incorporate collaboration. As previously discussed in Section 4.6 (i.e., Usage Scenarios), we recommend that researchers do not remove the Collaboration statements. The final version of the CSI handles situations in which collaboration is not applicable in that participants can mark these statements as "N/A." However, it is critical that collaboration still appear in the paired-factor comparisons, in order to keep one standardized version of the CSI. Standardization of the CSI is very important. By having a standardized survey that fits both collaborative and noncollaborative tools, it allows researchers to easily compare CSI scores on CSTs with collaboration and CSTs without collaboration. It also makes it easier to interpret research articles that report CSI scores, because the overall index score will have a uniform meaning. It is also the case that the paired-comparison section can be beneficial to noncollaboration studies. By asking participants whether collaboration was important to them for a particular creative task (i.e., in the paired-comparison section), researchers can understand whether collaboration is important to users, even if it was not studied or the tool does not support it. For all of these reasons, it is our recommendation that the Collaboration scale is not removed from the CSI.

Removing Any Scale. Similar to Collaboration, we do not recommend that researchers remove any scales on the CSI. The main purpose of the paired-factor comparison section of the CSI is to provide a weighted score. A weighted score means that factors that are less relevant will contribute less, through their lower weightings, to the CSI score. However, if a particular factor was problematic to a researcher for some reason, then the researcher may want to calculate the CSI with and without that factor and report both scores, but in this case the researcher should be very specific about how the modified score was calculated, and why it is being reported in addition to the standardized score.

Skipping the Paired-Factor Comparisons. In some cases, researchers may choose to not use the paired-factor comparisons. By not administering this section, it will be impossible for researchers to calculate CSI scores. Therefore, when reporting CSI results in a publication without paired-factor comparisons, researchers should explicitly state that they are reporting results from a particular scale(s) on the CSI, rather than the overall index score.

Rewording Questions. We strongly recommend against rewording any questions on the CSI. Not only will researchers not be able to report the CSI's overall index score (as in the previous scenario), but also researchers will also not be able to claim that they are using the standardized CSI. In the event that researchers should modify the way in which certain statements are worded, it will be especially important for researchers to calculate the reliability of their scales (i.e., Cronbach's alpha). In addition, it is also extremely important in these cases that researchers are explicit in their reporting of results that they are using statements that were *modified* from the CSI, and they should not report an index score. The CSI's index score is intended to be a standardized score that will be meaningful to other researchers, so it should be very clear to other researchers when the CSI itself is not being used.

8. FUTURE WORK

8.1. CSI Testing in Other Domains

In our future work, we plan to test the Creativity Support Index in user studies investigating creativity support tools outside of the arts domain. Specifically, we are interested in testing the CSI in software that supports creative programming, visual analytics, mathematics, and engineering design. In addition to testing the CSI in a variety of tools, we would also like to investigate the effectiveness of employing the CSI as an additional evaluation method for studying creativity support environments.

8.2. Relationship to Tool Complexity

Throughout the course of this research, we have continually noted a relationship between creativity support and tool complexity. As with productivity tools, there is a tradeoff between simplicity and power. A simple tool may be easy to learn and effective for simple tasks, but will likely not have the power to support the complex flow of tasks that is often inherent in creative work. On the other hand, a powerful tool that supports complex creative task flows may be difficult to learn and actually take years to master. Other researchers have noted the link between creativity support and skill level, and Csikszentmihalyi [1997] has identified the challenge level being met by skill level as characteristic of flow experiences. We believe that tool complexity and user skill level impact CSI scores. Our testing shows that more complex tools appear to score higher on the CSI than less complex tools, but this seems to be affected in part by the difficulty of the creative task. For example, photography students learning to master photo editing generated a CSI score of 84 for Adobe Photoshop, while sketching participants generated a CSI score for AutoDesk SketchBook Express of 64, but SketchBook Express is a very simplified sketching program. Alternatively, a custom color plug-in used for a coloring activity generated a CSI of 76, and while this tool is very simple, the task was also very simple. We plan to develop a system complexity metric that can help designers further investigate these relationships.

9. CONCLUSION

In this article, we have presented a psychometric tool called the Creativity Support Index (CSI), which is a survey metric designed for evaluating how well a tool supports a user doing creative work. The power of the CSI lies in the factorization of creativity: rather than trying to define creativity, we look at what factors are most relevant to supporting creative work processes. Additionally, by modeling the survey after the NASA TLX, we have gained the flexibility that comes with weighting factors by importance to the task. This means that the CSI can be used across a wide variety of open-ended creative tasks. We have provided examples that illustrate how to make use of factor scores when planning to improve or redesign Creativity Support Tools.

The CSI is a measurement contribution to the academic community in that it can assist researchers and developers working to design and refine creativity support tools. In concert with other evaluation metrics, the CSI can help ensure that CSTs are effectively supporting people in their creative work. We made the CSI available in the form of a desktop application, which researchers can use to administer and automatically score the CSI. In addition to presenting the CSI, we have detailed our entire development process. It is our expectation that this will be useful to other HCI researchers interested in developing psychometric tools.

APPENDIX

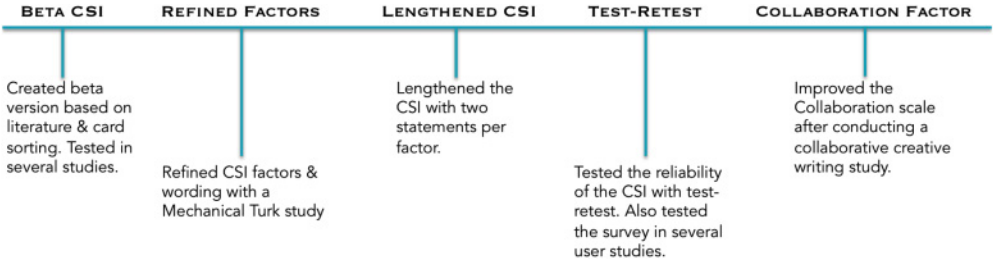
A. DEVELOPMENT OF THE CREATIVITY SUPPORT INDEX

The development of the CSI followed a rigorous, iterative process, which is the standard in the development of psychometric tools. In this appendix, we document our methodology, which will be beneficial to other researchers who are interested in developing psychometric tools for HCI. Figure 5 provides an overview of our development process.

A.1. Beta Version

We began developing the CSI by creating a beta version that was based on concepts from the literature on creativity and cognition support tools (Figure 6). Since many of the concepts overlapped, it made sense to synthesize them through a card sorting process. We grouped these concepts into six categories and then named each category, which became the six factors on the Beta CSI: Exploration, Collaboration, Engagement, Effort/Reward Tradeoff, Tool Transparency, and Expressiveness. We then wrote one agreement statement per factor for a total of six items (Figure 7).

A.1.1. Study 1: Ken Burns Study. The beta CSI was deployed in a within-subjects experiment (N = 32), in which we used the beta CSI to compare two different interaction techniques for specifying Ken Burns regions in the creation of photographic slideshows. One technique used two mice and two cursors to select the rectangular regions of interest in the photograph (similar to a cropping tool), while the other technique used



Created beta version based on literature & card sorting. Tested in several studies.

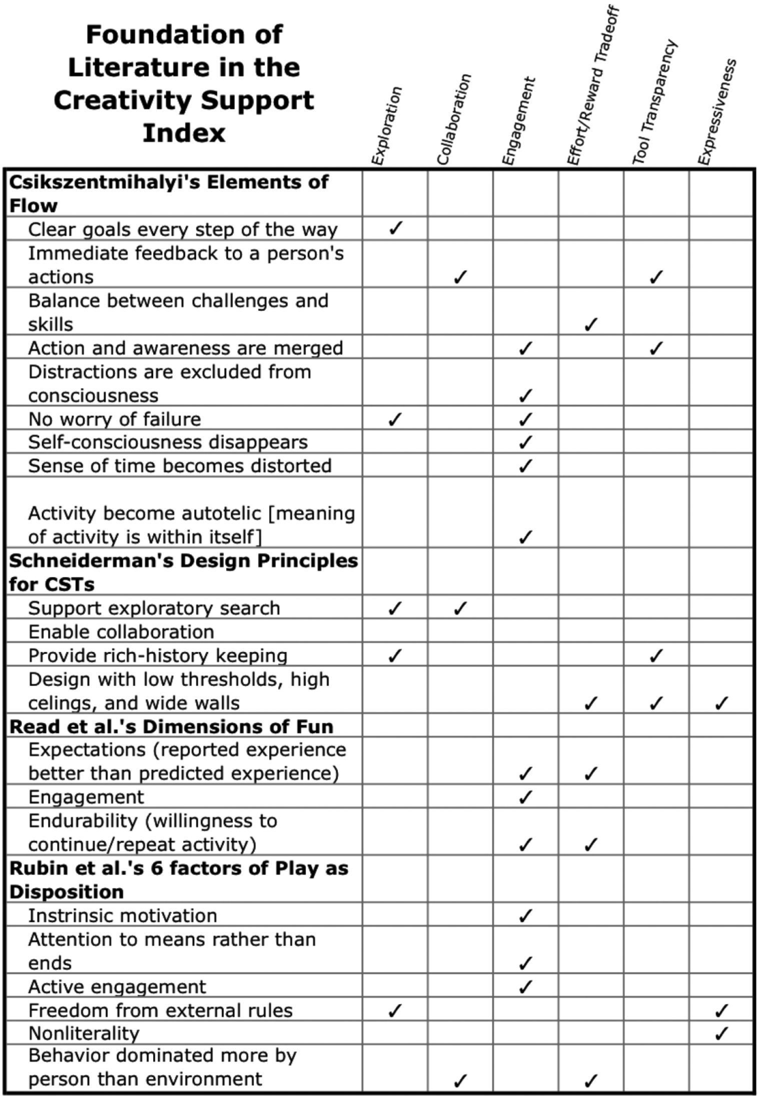
Refined CSI factors & wording with a Mechanical Turk study

Lengthened the CSI with two statements per factor.

Tested the reliability of the CSI with test-retest. Also tested the survey in several user studies.

Improved the Collaboration scale after conducting a collaborative creative writing study.

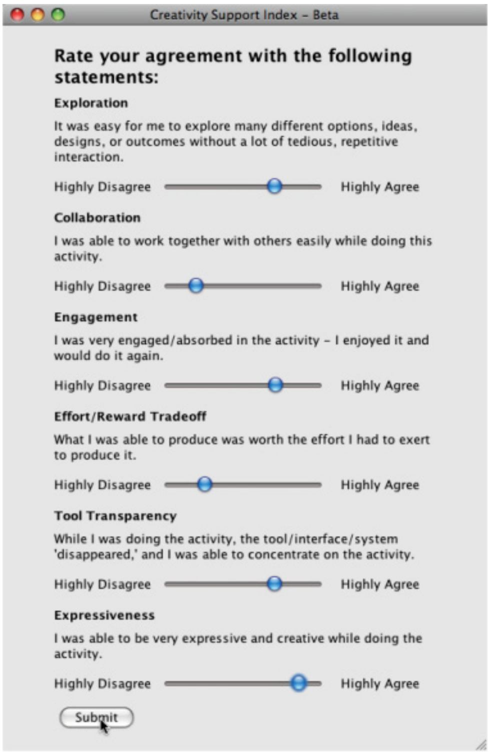
Fig. 5. A summary of the CSI's development process.



Foundation of Literature in the Creativity Support Index

	Exploration	Collaboration	Engagement	Effort/Reward Tradeoff	Tool Transparency	Expressiveness
Csikszentmihalyi's Elements of Flow						
Clear goals every step of the way	✓					
Immediate feedback to a person's actions		✓			✓	
Balance between challenges and skills				✓		
Action and awareness are merged			✓		✓	
Distractions are excluded from consciousness			✓			
No worry of failure	✓		✓			
Self-consciousness disappears			✓			
Sense of time becomes distorted			✓			
Activity become autotelic [meaning of activity is within itself]			✓			
Schneiderman's Design Principles for CSTs						
Support exploratory search	✓	✓				
Enable collaboration						
Provide rich-history keeping	✓				✓	
Design with low thresholds, high celings, and wide walls				✓	✓	✓
Read et al.'s Dimensions of Fun						
Expectations (reported experience better than predicted experience)			✓	✓		
Engagement			✓			
Endurability (willingness to continue/repeat activity)			✓	✓		
Rubin et al.'s 6 factors of Play as Disposition						
Instrinsic motivation			✓			
Attention to means rather than ends			✓			
Active engagement			✓			
Freedom from external rules	✓					✓
Nonliterality						✓
Behavior dominated more by person than environment		✓		✓		

Fig. 6. We present the relationships between the primary creativity theories and related concepts, which resulted in the six factors on the beta CSI.



Creativity Support Index – Beta

Rate your agreement with the following statements:

Exploration

It was easy for me to explore many different options, ideas, designs, or outcomes without a lot of tedious, repetitive interaction.

Highly Disagree — Highly Agree

Collaboration

I was able to work together with others easily while doing this activity.

Highly Disagree — Highly Agree

Engagement

I was very engaged/absorbed in the activity – I enjoyed it and would do it again.

Highly Disagree — Highly Agree

Effort/Reward Tradeoff

What I was able to produce was worth the effort I had to exert to produce it.

Highly Disagree — Highly Agree

Tool Transparency

While I was doing the activity, the tool/interface/system 'disappeared,' and I was able to concentrate on the activity.

Highly Disagree — Highly Agree

Expressiveness

I was able to be very expressive and creative while doing the activity.

Highly Disagree — Highly Agree

Fig. 7. This is an electronic version of the beta CSI, which was deployed in three user studies. The beta CSI also included a paired-factor comparison section (not pictured).

panning and zooming and was based on the Ken Burns Effect interface in Apple's iPhoto 2008.

Since this study did not involve collaboration, we expected to see low ratings on the Collaboration item. The Collaboration statement was, "I was able to work together with others easily while doing the activity," on a scale of "Highly Disagree" to "Highly Agree." However, only nine of 32 participants selected the lowest values. A few participants also verbally asked if they should ignore the item. Based on these results, we believed that this statement was phrased poorly since it focused on the collaborative nature of the *activity*, rather than on the collaborative affordance of the *tool*.

A.1.2. Study 2: Color Exploration Study. The beta CSI was also employed in a study for evaluating a bimanual color exploration tool with eight participants, who were digital artists, designers, or architects. Collaboration and Tool Transparency were both problematic factors in this study. Two of the eight participants wrote "N/A" beside the Collaboration statement. The statement for Tool Transparency was, "While I was doing the activity, the tool/interface/system 'disappeared,' and I was able to concentrate on the activity." In response to this item, one person wrote, "Yes, it disappeared, but it would have been easier if it stayed." This participant's comment indicated a clear issue with the wording of the Tool Transparency item, since it appears that this participant thought we were referring to a tool within the interface actually disappearing from view inside the application.

A.1.3. Study 3: Kinematic Templates Study. The beta CSI was used in a longitudinal study on a drawing program that made use of varying control-display ratios to allow for a variety of kinematic drawing effects [Fung et al. 2008]. Each participant was involved in four or five sessions over the course of 3 to 12 weeks, with each session lasting approximately 1 hour. After each session, participants were given a paper version of the CSI to complete. Similar to Study 1 and Study 2, participants were confused by the Collaboration item. After the first participant expressed confusion over this item, the beta CSI was altered to remove the Collaboration item, both from the ratings and the paired comparisons. Removing this item allowed us to explore using a different version of the CSI when collaboration was not relevant.

The item about exploration also brought up some concern in this study. The Exploration item was, “It was easy for me to explore many different options, ideas, designs, or outcomes without a lot of tedious, repetitive interaction.” In response to this item, one participant said, “I kind of like tedious, repetitive interaction. . . it’s just the way I draw.” This participant was observed using the same action or template repeatedly to draw specific features but not in exploring different alternatives. This feedback indicated that this statement needed rewording.

A.2. Refining CSI Factors: Creativity Word Study

The beta CSI studies helped to identify several issues. We were concerned that people may use different terminologies than those used in the creativity literature and subsequently in the beta CSI. We also wanted to be more thorough in our creativity factor categorization than our card sorting process may have allowed. Therefore, we developed a word study following a process similar to the early NASA TLX development, in which the authors presented people with a list of words potentially related to workload and then asked them how much each word was related to workload [Hart and Staveland 1988].

In this study, 300 people from Amazon Mechanical Turk rated 19 words according to how much they were related to the creative process on a scale of “Extremely Important” to “Not At All Important.” These words were selected from creativity research and from common parlance when describing the creative process. The words and their ratings are shown in Figure 8. Following a process similar to Hart and Staveland [1988], we used a principal components analysis with an extraction method based on components that had an eigenvalue above 1.0 [Kaiser 1960], which extracted six components. We then named the six extracted components based on the words with the heaviest load in each component. These became the six factors for the final CSI and are available in Table VIII. We excluded the component that had “motivation” and “imagination” as the heavy loading words because the CSI is a measurement tool that focuses on the *tool*, rather than *person*.

In Table VIII, participants rated almost all of these 19 words as important to the creative process, except for collaboration. Only 35.60% of participants said that collaboration was essential to the creative process (Figure 8). As a result, collaboration did not load strongly on any factor. We believe that collaboration did not reflect highly in these ratings because a very strong stereotype exists that the creative genius works alone [Howe 2000]. However, even though collaboration was not rated as important to creativity as other factors were, we strongly agree with the creativity research on the importance of collaboration [Howe 2000; Shneiderman 2007]. Therefore, we kept collaboration as a factor on the CSI.

After this study, we arrived at the factors (or scales) that are now used in the final version of the CSI: Results Worth Effort, Expressiveness, Exploration, Immersion, Enjoyment, and Collaboration. Table VIII specifies how these new factors related to the original names in the beta CSI.

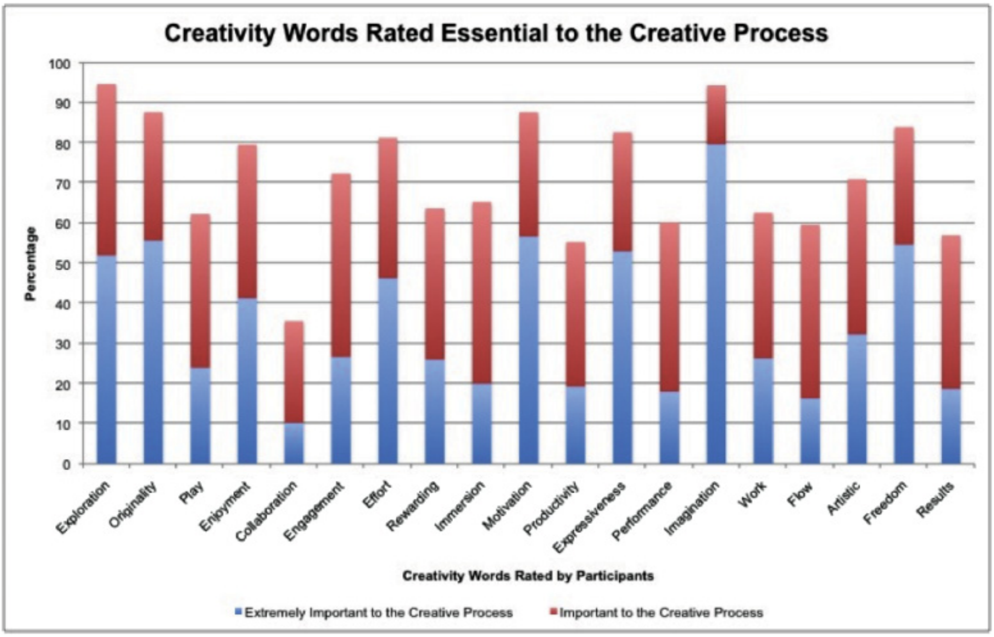


Fig. 8. Three hundred participants rated how much each of these 19 words were related to the creative process. The purpose of this study was to refine the factors used on the CSI.

Table VIII. The Mechanical Turk Study on Creativity Word Ratings Resulted in New Factors for the CSI Based on How These Words Loaded on Components, As a Result of a Principal Components Analysis

The bold-faced components were weighted the heaviest.

Extracted Components From Creativity Words	CSI	Beta CSI
collaboration, effort, **work**, **productivity**, **performance**, rewarding, **results**	Results Worth Effort	Effort/Reward Tradeoff
play, enjoyment, flow, **expressiveness**, freedom, **artistic**	Expressiveness	Expressiveness
effort, **motivation**, **imagination**	*n/a*	*n/a*
exploration, play, engagement, collaboration	Exploration	Exploration
flow, **immersion**	Immersion	Tool Transparency
enjoyment	Enjoyment	Engagement
	Collaboration	Collaboration

A.3. Lengthening the CSI

An initial goal of the CSI was for it to be similar in form and length to the NASA TLX [Hart and Staveland 1988]. However, based on best practices in psychometric research, we realized that it was imperative to have multiple statements for each factor. With only one item, it would be impossible to treat each factor as its own scale, which means that we would not be able to compute the reliability (e.g. Cronbach's alpha) of each factor. Being able to calculate the reliability of a scale not only is important when developing a psychometric tool but also is important to researchers who want to calculate their own reliability ratings, as relevant to their user study. To address reliability calculations, we lengthened the CSI to have two agreement statements (or items) per factor. While even more items per factor could have further increased the reliability, we believed it was important that research participants be able to fill out the CSI quickly.

In order to lengthen the CSI, we created an extended version of the survey by writing additional agreement statements. In this extended version, there were 47 total statements with seven to eight statements for every factor. It is common practice in psychometrics to create longer, temporary versions of a survey when developing a new metric. This allows the researcher to conduct an item-level analysis, allowing identification of the items that perform the best. In our case, we only needed the top two performing statements for each of the six factors. The factors used were the finalized factors from the Mechanical Turk study that was described in Section A.2. It is important to note that we only deployed the agreement statements and did not deploy the paired-factor comparisons since the purpose of the study was to find the best agreement statements.

We deployed this temporary, extended version of the CSI to 70 participants using six different CSTs. Specifically, 17 people used Adobe Flash, two used Adobe Illustrator, five used Adobe InDesign, eight used Adobe Photoshop, four used a combination of InDesign and Photoshop, 14 used Apple Final Cut Pro, and 20 used a custom tool called LayerCake. Typically, the CSI was deployed in a classroom or workshop setting, where the participants had tasks to complete, so we did not give participants any specific tasks. Instead, we allowed them to continue working on the activities that they were presently engaged in. We instructed them to fill out the CSI after they were finished using their creativity support tool. The advantage of this method is that we collected data across a variety of tasks, for a variety of tools, which makes our results more generalizable. It is important to note, however, that none of these CSTs were collaborative, although it is possible that some students may have collaborated on their tasks outside of the tool.

We used a factor analysis to analyze the data, which is similar to a principal components analysis but is the preferred method in psychometrics for understanding the relationship between survey items. In the factor analysis, we used the Kaiser rule [1960] for our extraction method, and we selected an oblique rotation because we had no reason to believe that our factors were not correlated. We began by running factor analyses on survey items for each of the six scales. For example, we performed a factor analysis on the seven items that we wrote for the Results Worth Effort scale, a factor analysis on the eight items for the Exploration scale, and so on. The purpose of running a factor analysis on the items in each individual scale was to verify that the items in each scale only represented one factor. We found that three of our scales (Collaboration, Immersion, and Expressiveness) had items that loaded heavily (above 0.40) on two factors. The items that loaded onto multiple factors could be problematic in that they may actually be measuring a different construct than we intended to measure; therefore, we discarded any items that loaded on multiple factors.

After using a factor analysis to verify one factor per scale, we performed an item-level analysis to assess the reliability of each scale and to also reduce each scale to two statements, as previously discussed. In calculating the reliability of each scale, we immediately removed any items that had a corrected item-total correlation (CITC) of $r < 0.60$. Items with a low CITC may indicate that the item is not measuring what the rest of the scale is measuring. While anything above 0.40 would be acceptable, we were able to set higher standards because we only wanted the two best-performing questions for each scale. After reducing each scale using this criterion, we eliminated items that did not seem to fit with the context of the scale and also items that had very similar wording. Any remaining elimination was done based on CITC. The survey items and Cronbach's alphas for each scale are available in Table IX.

Table IX. We lengthened the Creativity Support Index to Have Two Items per Factor
In this table, we also report the reliability (or Cronbach's alpha) of each scale.

Results Worth Effort 1. What I was able to produce was worth the effort I had to exert to produce it. 2. I was satisfied with what I got out of the system or tool.	Alpha: .925
Exploration 1. The system was helpful in allowing me to track different ideas, outcomes, or possibilities 2. It was easy for me to explore many different ideas, options, designs, or outcomes.	Alpha: .734
Collaboration 1. It was really easy to share ideas and designs with other people inside this tool. 2. The system or tool offered support for multiple users.	Alpha: .827
Immersion 1. I became so absorbed in the activity that I forgot about the system or tool that I was using. 2. My attention was fully tuned to the activity, and I forgot about the system or tool that I was using.	Alpha: .707
Expressiveness 1. I felt very artistic while using this system or tool. 2. I was able to be very creative while doing the activity.	Alpha: .900
Enjoyment 1. I would be happy to use this system or tool on a regular basis. 2. I enjoyed this system or tool.	Alpha: .930

ACKNOWLEDGMENTS

We would like to thank Dr. Michael Terry and Richard Fung for their support during the early phases of CSI development. We also would like to thank the various researchers, such as Berto Gonzalez, who used the tool and provided feedback. We thank Dr. Eric Heggestad for guidance on the psychometric process. We thank Dr. Mary Lou Maher, Dr. Heather Lipford, and Dr. Scott Klemmer for their insightful feedback. We thank Jordan Bullington for her assistance with the collaborative creative writing study. Finally, we thank the members of the UNC Charlotte HCILab for their feedback and support of this work.

REFERENCES

Robert S. Albert and Mark A. Runco. 1999. A history of research on creativity. In *Handbook of Creativity*. Springer.

Teresa M. Amabile. 1996. *Creativity in Context: Update to the Social Psychology of Creativity*. Westview Press.

Ronald A. Beghetto and James C. Kaufman. 2007. Toward a broader conception of creativity: A case for "mini-c" creativity. *Psychology of Aesthetics, Creativity, and the Arts* 1, 2, 73–79.

Margaret A. Boden. 2004. *The Creative Mind: Myths and Mechanisms*. Psychology Press.

Linda Candy. 2013. Evaluating creativity. In *Creativity and Rationale: Enhancing Human Experience by Design*. Springer.

Erin A. Carroll and Celine Latulipe. 2009. The creativity support index. In *ACM CHI 2008 Extended Abstracts*.

Erin A. Carroll and Celine Latulipe. 2012. Triangulating the personal creative experience: Self-Report, external judgments, and physiology. In *Proceedings of Graphics Interface 2012*.

Erin A. Carroll, Celine Latulipe, Richard Fung, and Michael Terry. 2009. Creativity factor evaluation: Towards a standardized survey metric for creativity support. In *Proceedings of ACM Creativity & Cognition 2009*.

Mihaly Csikszentmihalyi. 1997. *Creativity: Flow and the Psychology of Discovery and Invention*. Harper Perennial.

Richard Fung, E. Lank, Michael Terry, and Celine Latulipe. 2008. Kinematic templates: End-user tools for content-relative cursor manipulations. In *Proceedings of ACM UIST'08*.

Berto Gonzalez and Celine Latulipe. 2011. BiCEP: Bimanual color exploration plugin. In *ACM CHI'11 Extended Abstracts*.

J. P. Guilford. 2001. *Creative Projects with LEGO Mindstorms*. Addison-Wesley Professional.

Sandra G. Hart. 2006. NASA task load index (NASA TLX): 20 Years Later. In *Human Factors and Ergonomics Society Annual Meeting Proceedings*.

Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Human Mental Workload* (1988).

Tom Hewett, Mary Czerwinkski, Michael Terry, Jay Nunamaker, Linda Candy, Bill Kulers, and Elisabeth Sylvan. 2005. Creativity support tool evaluation methods and metrics. In *NSF Workshop Report on Creativity Support Tools*.

Dennis Hocevar. 1979. Ideational fluency as a confounding factor in the measurement of originality. *Journal of Educational Psychology* 71, 2, 191–196.

Dennis Hocevar and Patricia Bachelor. 1989. A taxonomy and critique of measurements used in the study of creativity. In *Handbook of Creativity*. Springer.

Michael J. A. Howe. 2000. *Genius Explained*. Cambridge University Press.

H. F. Kaiser. 1960. The application of electronic computers to factor analysis. *Educational and Psychological Measurement* 20, 141–151.

James C. Kaufman and Ronald A. Beghetto. 2009. Beyond big and little: The four c model of creativity. *Review of General Psychology* 13, 1, 1–12.

Andruid Kerne, Steven M. Smith, Eunyee Koh, Hyun Choi, and Ross Graeber. 2008. An Experimental Method for Measuring the Emergence of New Ideas in Information Discovery. *International Journal of Human-Computer Interaction* 24(5), 460–477.

M. J. Kim and Mary Lou Maher. 2005. *Comparison of Designers Using a Tangible User Interface and a Graphical User Interface and the Impact on Spatial Cognition*. Technical Report.

Celine Latulipe. 2013. The value of research in creativity and the arts. In *Proceedings of ACM Creativity and Cognition'13*.

Celine Latulipe, Ian Bell, C. Clarke, and C. Kaplan. 2006. symTone: Two-handed manipulation of tone reproduction curves. In *Proceedings of GI'06*.

J. Read, S. MacFarlane, and C. Casey. 2002. Endurability, engagement, and expectations: Measuring children's fun. In *IDC'02*.

Ben Shneiderman. 2007. Creativity support tools: Accelerating discovery and innovation. *Commun. ACM* 50, 12, 20–32.

B. Shneiderman, G. Fischer, and M. Czerwinski. 2006. Creativity support tools: Report from a US National Science Foundation sponsored workshop. *International Journal of Human-Computer Interaction* 20, 2, 61–77.

Received November 2012; revised July 2013, February and April 2014; accepted May 2014