

QUANTIFYING THE PERSONAL CREATIVE EXPERIENCE: EVALUATION
OF DIGITAL CREATIVITY SUPPORT TOOLS USING SELF-REPORT AND
PHYSIOLOGICAL RESPONSES

by

Erin Ashley Carroll

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2013

Approved by:

Dr. Celine Latulipe

Dr. Heather Lipford

Dr. Mary Lou Maher

Dr. Eric Heggstad

Dr. Michael Terry

©2013
Erin Ashley Carroll
ALL RIGHTS RESERVED

ABSTRACT

ERIN ASHLEY CARROLL. Quantifying the personal creative experience: evaluation of digital creativity support tools using self-report and physiological responses. (Under the direction of DR. CELINE LATULIPE)

Creativity is understood intuitively, but it is not easily defined and therefore difficult to measure. This makes it challenging to evaluate the ability of a digital tool to support the creative process. When evaluating creativity support tools (CSTs), it is critical to look beyond traditional time, error, and other productivity measurements that are commonly used in Human-Computer Interaction (HCI) because these measures do not capture all the relevant dimensions of creativity support. Unfortunately, there are no clear measures of success to quantify in regards to creativity support tools, and this lack of ‘convenient’ metrics is a real challenge to their evaluation.

In this dissertation, I introduce two computational methodologies for evaluating creativity support tools, including: (1) the Creativity Support Index (CSI), which is a psychometrically developed and validated survey, designed for evaluating the ability of a tool to support the creative process of users, and (2) a novel sensor data approach to measuring ‘in-the-moment-creativity’ (ITMC), to detect moments when users experience high creativity using electroencephalography (EEG), activity metrics (e.g., keyboard/mouse logger and accelerometer data), and machine learning.

ACKNOWLEDGEMENTS

There have been many people that helped to make this dissertation possible, and more generally, helped me to succeed in graduate school. I first want to sincerely thank my advisor, Dr. Celine Latulipe. I owe much of my success in graduate school to her mentorship. It is also because of Celine that I entered this PhD program: she talked with me about graduate school back in Fall 2007, when I was an undergraduate student in her HCI class. Celine has been an extremely supportive advisor in every capacity, including mentoring me on writing research papers, giving conference talks, and preparing me to teach. I feel very privileged to be the first PhD student that trained under her.

I would also like to thank my dissertation committee: Dr. Heather Lipford, Dr. Mary Lou Maher, Dr. Eric Heggstad, and Dr. Michael Terry. First, Dr. Heather Lipford has been a mentor to me throughout graduate school, and it was Heather that gave me my first opportunity to be a Research Assistant when I was a senior in undergrad. I would also like to thank Heather for helping me navigate the job market, and I also enjoyed being her Teaching Assistant in the last semester of graduate school. I am extremely honored to have Dr. Mary Lou Maher serve on my committee. I am very grateful for the time that she spent giving me constructive feedback on my dissertation, as well as the insights that she provided during my dissertation defense. The committee member that I have known the longest is Dr. Eric Heggstad, who was my professor while completing my undergraduate degree in the Psychology department. Eric's assistance in psychometrics was instrumental to me developing

the Creativity Support Index. Finally, I would like to thank Dr. Michael Terry at the University of Waterloo. Over the years, Michael has provided me feedback on my research, including my first year of graduate school when he collaborated with Celine and I on the first Creativity Support Index paper. In addition to his contributions to my research, I would also like to thank Michael for participating in my dissertation defense remotely while on sabbatical in the United Kingdom.

There are other professors at UNC Charlotte that I would like to thank. Dr. David Wilson served as a mentor to me throughout graduate school as part of the Dance.Draw project, and he also served on my dissertation proposal committee. I also want to extend a very sincere thank you to Dr. Richard Souvenir for his machine learning consultations that he provided me. I am extremely grateful to Richard for his expertise, patience, and even letting me cry in his office. Finally, I want to thank Dr. Teresa Dahlberg for awarding me with the GAANN Fellowship¹, which not only funded my PhD but also paid for my conference travel and the equipment that I utilized to complete my dissertation.

Completing this dissertation would certainly not be possible without the friendships that I forged in graduate school. I want to sincerely thank all of my colleagues in the HCI Lab, including Berto Gonzalez, Vikash Singh, Andrew Besmer, and Jason Watson. In particular, I must thank my academic siblings, Berto and Vikash, for always being there for me, both as a friend and as a fellow academic. I would also like to acknowledge my other collaborators on the Dance.Draw project: Sybil Huskey, Melissa Word, and Dr. Danielle Lottridge. Collaborating with these women

¹GAANN: Graduate Assistance in Areas of National Need, US Department of Education

has also made a contribution to my research throughout graduate school. In particular, I would like to thank Danielle for her mentorship in writing conference papers. I am extremely grateful for the professional advice that was provided to me by PhD students that graduated before me: Dr. Evan Suma, Dr. Pamela Wisniewski, and Dr. Jing Xiao.

I would also like to sincerely thank my family: my mom and stepfather, Renee and Chad McDaniel; my sister Christa Henrickson; and my grandparents, Robert and Patricia Morrison. My family's support and encouragement throughout my life has been instrumental to my success. Finally, I would like to thank my fiancé, Jason Cherry, for loving me.

TABLE OF CONTENTS

LIST OF FIGURES	ix
LIST OF TABLES	x
CHAPTER 1: INTRODUCTION	1
1.1 Creativity Support Tools	3
1.2 Contributions	7
1.3 Dissertation Overview	8
CHAPTER 2: CREATIVITY	10
2.1 Creativity Theory	10
2.2 Concepts Relevant to Creativity	14
2.3 Creativity Measurement Approaches	15
CHAPTER 3: THE CREATIVITY SUPPORT INDEX (CSI)	20
3.1 The Creativity Support Index	21
3.2 Development of the Creativity Support Index	34
3.3 Test-Retest Reliability	46
3.4 User Studies with the CSI	49
3.5 Collaboration	52
3.6 Discussion	56
3.7 Future Work	60
CHAPTER 4: IN-THE-MOMENT-CREATIVITY (ITMC): STUDY 1	62
4.1 Background: Physiological Metrics for Creativity	63
4.2 ITMC Measurement Approach	66

4.3	ITMC Study 1	67
4.4	Hypotheses	79
4.5	Results	80
4.6	Discussion	84
CHAPTER 5: EXTENDING ITMC MEASUREMENT: STUDY 2		88
5.1	ITMC Application Improvements	88
5.2	ITMC Study 2 Methodology	91
5.3	Data Processing	97
5.4	Machine Learning Procedure	100
5.5	Classification Results	101
5.6	Video Analysis	110
5.7	Discussion	111
5.8	Future Work	112
CHAPTER 6: CONCLUSION		113
6.1	Dissertation Contributions	113
6.2	Future Research	116
REFERENCES		119
APPENDIX A: CREATIVITY SUPPORT INDEX		124
APPENDIX B: MATERIALS FROM ITMC		131

LIST OF FIGURES

FIGURE 1: Novelty-Impact space of creativity	12
FIGURE 2: Final CSI software: 6 out of 12 agreement statements	25
FIGURE 3: Final CSI Software: 2 of the 15 Paired-Factor Comparisons	25
FIGURE 4: Equation for scoring the CSI	27
FIGURE 5: Beta CSI: Card sorting of creativity literature	33
FIGURE 6: Beta CSI: Screenshot of agreement statements	35
FIGURE 7: Beta CSI: Longitudinal CSI results	39
FIGURE 8: Creativity word ratings from Mechanical Turk study	40
FIGURE 9: Photographs from creative writing study on Google Docs	55
FIGURE 10: ITMC triangulation	67
FIGURE 11: ITMC reporting application	70
FIGURE 12: Photos of Emotiv EEG and Affectiva Q	72
FIGURE 13: ITMC reporting app: Chunking problem	89
FIGURE 14: Griffin PowerMate 3D mouse	91
FIGURE 15: ITMC 2.0 reporting application	92
FIGURE 16: ITMC 2.0: Rewind menu	93
FIGURE 17: ITMC Study 2: Caption example	96
FIGURE 18: ITMC Study 2: Image example	98
FIGURE 19: Sketching correlation results	104
FIGURE 20: Writing correlation results	104
FIGURE 21: Classifier comparison across sensors	105

LIST OF TABLES

TABLE 1: Overview of creativity support tools	5
TABLE 2: Final version of the CSI: 12 agreement statements	24
TABLE 3: Final version of the CSI: 15 paired comparisons	24
TABLE 4: Sample reporting of CSI results	28
TABLE 5: Final CSI factors from the Mechanical Turk study	43
TABLE 6: Lengthened CSI (i.e. two items per factor)	47
TABLE 7: CSI test-retest results	48
TABLE 8: Summary of user studies with the CSI	51
TABLE 9: Collaboration scale	56
TABLE 10: Summary of post experiment surveys	83
TABLE 11: Comparison of ITMC Study 1 and Study 2	94
TABLE 12: Classifier performance across sensors	106
TABLE 13: Classifier performance for cross-task generalizability	107
TABLE 14: Post experiment survey results	109
TABLE 15: ITMC Study 2: Sketching correlations	131
TABLE 16: ITMC Study 2: Writing correlations	132

CHAPTER 1: INTRODUCTION

Creativity is revered by society as an essential human resource: it is linked to the economic principles of innovation, productivity, and growth, as well as to more personal concepts of creative expression, do-it-yourself (DIY) culture, and psychological flourishing. Creativity is deeply rooted in our culture to the extent that most individuals strive to be more creative and unique. In this capacity, digital creativity support tools (CSTs) have the potential to make substantial impact on both individuals and society by offering support for scientific, engineering, humanist, and artistic endeavors. CSTs are able to influence creative work that happens every day, in addition to helping bring about those rare creative moments that can change history and have vast impacts on society. In addition to supporting and influencing creative work, CSTs also have the potential to help people on their own personal, creative journeys.

There is a substantial body of work in HCI that guides the evaluation of productivity support tools. Shneiderman compared the growing community of researchers developing and studying creativity support tools to the earlier rise of researchers working on productivity support tools [57]. He said that researchers in CSTs are “moving from the comparatively safe territory of productivity support tools to the more risky frontier of creativity support tools.” Shneiderman noted that one of the challenges that makes CST research ‘risky’ is that there are no obvious measures of success [57]. In productivity support tools, researchers often employ time and error

measurements as evaluation metrics. However, these metrics are not always appropriate to the evaluation of CSTs. For example, if a user was to spend a long time performing a task with a productivity support tool, this may indicate that the tool was deficient. In contrast, if a user was to spend a long time working with a CST, this may actually indicate deep engagement, rather than a tool deficiency. This is because a user may have become immersed in their creative work and lost track of time. When tasks are open-ended, like many creative activities, time is not an appropriate metric. Even when time and error metrics can successfully be employed to study CSTs, they do not usually tell the complete story. For example, when Latulipe et al. created the symSpline technique [43], a standard formal experiment showed that users could complete spline matching tasks significantly faster using the dual cursor technique. However, the time metric used in that experiment was not sufficient to demonstrate the fluid expressiveness and ease of exploration afforded by the symSpline technique.

It is also the case that ‘errors’ are not always applicable to CST evaluation. Artists and creative individuals often view mistakes as opportunities for serendipitous discovery, rather than errors. Because of the open-ended nature of creative work, random or pseudo-random inputs can often be helpful during certain phases of work by providing juxtaposing concepts, new ways of envisioning artifacts, or unexpected variations on designs. These types of serendipitous, fortuitous errors may help with unblocking design fixations, a well-documented problem in creative work [18]. Thus, in the world of productivity evaluation, researchers may want to study the ‘wrong paths’ that a user takes in trying to accomplish a task, in order to redesign the tool and prevent others from taking those wrong paths in the future. The opposite is likely true for

creativity support tools, where exploration is fundamentally important.

In 2005, a report from an NSF workshop on creativity support tools recognized that better evaluation metrics were needed [28, 58]. Without effective evaluation methods for CSTs, designers are left to guess at how to improve their tools to better support users’ creative endeavors. From this NSF workshop, Hewett et al. [28] noted that CST evaluation should take a strong mixed-methods approach, and Shneiderman [58] called for better evaluation metrics to be developed. Since this workshop took place in 2005, there has been a heightened interest in evaluating CSTs and a number of researchers are developing a variety of evaluation methodologies [36]. These methodologies will be described later in Section 1.1.2.

In this dissertation, I introduce computational methodologies for evaluating creativity support tools. These approaches include a psychometric tool called the Creativity Support Index (or CSI) and a physiological sensor data approach called ‘in-the-moment-creativity’ (ITMC). The CSI is a user-based survey metric for evaluating creativity support tools, and it should be administered as a post-experiment survey. ITMC measurement is a novel approach for detecting when a user experiences high creativity, and it involves measuring physiological responses and using machine learning as a mechanism for classifying high creative experience.

1.1 Creativity Support Tools

1.1.1 Overview of Tools

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 purely within a collaborative environment. This dissertation is relevant to the larger class of creativity support environments, but it is primarily focused on creativity support tools.

CSTs span many different domains, and they support a variety of open-ended creative activities, including tools designed to support programming to information exploration to data analysis to artistic creations in visual, performing, or musical arts. In Table 1, I have compiled a summary of creativity support tools, including categories and examples of each, which builds upon Shneiderman’s summary [57]. Many of these tools can arguably be considered both a creativity support tool and a productivity support tool. This makes sense, as these tools often support a user across various iterative phases of a creative process: ideation, execution, and evaluation [8]. In fact, an individual or team may use several of these tools simultaneously or in sequence during a creative work process. The broader research looking at CSEs seeks to investigate the integration issues inherent in multiple tool usage, which is important. However, just looking at a single tool, one can see that productivity and creativity are mutually enforcing. Given the complexity of the iterative creative process, designers and researchers in CSTs must consider measures of success for both productivity and creativity.

Table 1: A summary of creativity support tools, including examples from research and industry. This table builds upon Shneiderman’s summary [57].

Category	Example
Visualization & Simulation	Tableau, D3, netLogo
Concept Mapping & Information Collage	combinFormation, Visio, Omnigraffle
Architectural & Design	AutoCAD, Rhino3D
Mathematics	SPSS, MatLab, WolframAlpha
Software development environments	Eclipse, Visual Studio
Video Editing	Final Cut Pro, iMovie
Drawing/Painting	Illustrator, InkScape, CorelDraw
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

1.1.2 Evaluation of Creativity Support Tools

While there is an extensive history of evaluating creativity, the evaluation of tools to support creativity is a much newer field of study. As previously discussed, Shneiderman noted that the evaluation of creativity support tools is challenging because there are no obvious metrics for creativity [57]. In addition, CST evaluation is also challenging because creativity is multi-dimensional, and therefore, a particular CST may not be intended to support all parts of the creative process. Hewett et al. concluded that there is no one-size-fits-all approach to evaluating CSTs and that one metric alone will not be enough [28]. Therefore, it is essential that researchers have a variety of metrics at their disposal, or in their ‘research toolbox.’

Various researchers have done one-off qualitative evaluations on a particular creativity support tool [60]. This often involves studying the tool’s use in the wild, interviewing the users, developing an understanding of the domain, and then recommending design changes based on the information gathered. This is an excellent approach, but it is time-consuming and does not easily allow for comparisons with similar creativity support tools.

While Shneiderman recognized the challenge of evaluating CSTs, he offered important design guidelines for CST researchers. He said that CSTs should: (1) Support exploratory search; (2) Enable collaboration; (3) Provide rich-history keeping; and (4) Design with ‘low thresholds, high ceilings, and wide walls.’ This latter idea can be related to the concept of freeform play, in which there are no rules or where rules can be made up but are malleable [7, 56]. It also suggests the need for tools that are simple to begin using, yet not overly simplistic, such that they have enough room for growth in terms of both users’ skill development and the scope of possible outcomes afforded by the tool.

Several HCI researchers have been working towards contributing better methodologies for evaluating creativity support tools, and as such, they are interested in putting together a suite of evaluation methodologies and metrics to be used by the community of CST researchers [36]. Some of these approaches focus on evaluating the outcomes created by people using the tools. Kerne et al. contributed evaluation metrics for studying information-based ideation tasks [37, 38]. These ideation tasks involve creative innovation, in which users’ goals are to develop new ideas, which is highly related to divergent thinking (an important dimension of creativity). Kerne et al.’s evaluation approach, called the Emergence Metric, involves judging the quality of synthesis in the generation of ideas, their novelty, their relevance, and counting the number of ideas created (i.e. fluency). A different approach is to study people while they are using a CST. For example, Kim and Maher developed a protocol analysis method for evaluating collaborative design and comparing graphical and tangible user interfaces [39].

My dissertation makes a contribution to the community of creativity support tool researchers by introducing computational methods for evaluating CSTs. These contributions are summarized in the following section.

1.2 Contributions

1. A psychometric tool called the Creativity Support Index (or CSI), which is a post-experiment survey designed for evaluating the ability of a system or tool to support the creative process. An electronic version of the CSI has been made available to the research community.
2. An application for the temporal, self-reporting of the creative work process. This application allows participants to report their ‘in-the-moment creativity’ (ITMC) ratings using a custom application. In this application, users are able to watch a screen recording video of their session using a creativity support tool and to rate time periods in which they felt creative.
3. A physiological sensor data approach to detecting moments of high creative experience using machine learning. The sensor data employed in this work includes electroencephalography (EEG), electrodermal activity (EDA), and activity metrics (i.e. keyboard/mouse logger, accelerometers, gyroscope).
4. Finally, my work in measuring ITMC sets the stage for more extensive research. By automatically detecting when a person is experiencing high creativity, we can: (1) Research which software features in a CST lead users into or out of creative experience peaks; (2) Develop adaptive interfaces that help people sustain creative experience peaks for a longer time; and (3) Quantify environmental or

contextual impacts on the creative experience when using a CST. Thus, successfully measuring ITMC will help researchers design and develop tools that promote and sustain periods of high creative experience.

1.3 Dissertation Overview

- Chapter 2 provides an overview of creativity research, including creativity theories, how creativity is measured, and concepts that are relevant to creativity and to creativity support tools.
- Chapter 3 presents the Creativity Support Index (CSI), as well as a discussion of its development and validation [10, 13]. The CSI is a psychometric tool designed for evaluating creativity support tools. It is intended to be used by researchers who are studying CSTs. Specifically, it should be administered as a post-experiment survey. There are six factors on the CSI, which include: Collaboration, Enjoyment, Exploration, Expressiveness, Immersion, and Results Worth Effort. The CSI is grounded in literature on creativity; concepts related to creativity; and creativity support tools.
- Chapter 4 introduces the concept of ‘in-the-moment-creativity’ (ITMC) and a computational approach to measure it [11]. In this study, ITMC measurement involves the triangulation of three temporal metrics: self-report ratings of creativity, external judgments of creativity, and physiological measurements through electroencephalography (EEG). Literature on creativity and physiology is also discussed in this chapter.

- Chapter 5 includes a second user study on ITMC. In this approach to ITMC measurement, I investigate whether moments of high creativity can be detected in a repeated measures study involving both sketching and writing. In addition to EEG, I also employ electrodermal activity (EDA) and activity metrics (i.e. keyboard/mouse logger, accelerometers, gyroscope). A new version of the ITMC application was also created based on observations from the ITMC study in Chapter 4.
- Chapter 6 provides a summary of the contributions of this dissertation, along with a discussion of future research directions.

CHAPTER 2: CREATIVITY

In this chapter, I summarize the creativity theory that is relevant to the evaluation of creativity support tools. This chapter is organized as follows: I begin with a discussion of creativity theory, defining creativity through the Novelty-Impact space, and then I present concepts that are relevant to creativity, such as play [7, 55, 56] and Csikszentmihalyi’s flow [15]. Finally, I end by discussing common approaches to measuring creativity, in which I emphasize the psychometric approach to creativity measurement.

2.1 Creativity Theory

Creativity has been studied extensively by researchers for decades, yet its complexity still makes it difficult to define and measure. While most people have an intuitive understanding of creativity, it is very difficult to define and therefore even more challenging to measure. Hewett et al. [28] presented a meta-analysis of psychological creativity research at the NSF workshop on creativity support tools [58], and concluded that: “Creativity can be considered to be the development of a novel product that has some value to the individual or to a social group.” However, these authors noted that psychology research does not clarify this definition of creativity any further.

Within psychology, creativity has been studied extensively using a variety of approaches with rigorous methodologies. While all creativity theories or definitions include the characteristic of *novelty* as being a component of creativity, the *impact* dimension also appears to be critical. In other words, the measure of whether something is creative depends on its novelty, as well as the contribution’s impact on society. For example, creativity is sometimes divided into Historical Creativity (H-creativity) or Psychological Creativity (P-creativity). In H-creativity, a contribution is creative if something is novel for the first time in history [7]. In contrast, P-creativity specifies that a contribution is creative if it is novel to the person that came up with the idea. Therefore, P-creativity is very common, whereas H-creativity is extremely rare. This bisection of creativity seems overly simplistic, and in fact, other researchers have approached creative contribution classification with finer granularity and with an eye to other relevant dimensions. For example, Maher has worked on computational methods for creativity evaluation that include representations of the element of surprise, which relates to how existing artifacts frame expectations for what might come next [45].

A different approach to classification of creative contributions is a two-dimensional approach that takes into account both novelty and the contribution’s impact on society. This stems from the work of Csikszentmihalyi [15], who investigated ‘Big-C creativity’ by studying highly eminent people that made large contributions to society: Pulitzer prize winners, professional athletes, renowned artists, etc. This led to other researchers identifying little-c creativity, mini-c creativity, and Pro-C creativity [6, 35]. These categories are represented in the Novelty-Impact space in Figure 1.

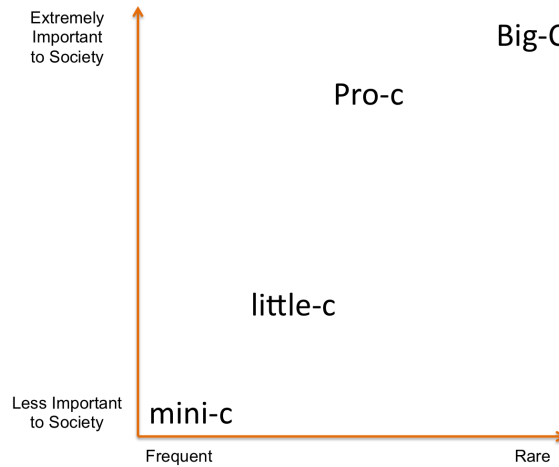


Figure 1: The creativity literature contains classifications of creative contributions across two dimensions: the Novelty-Impact space. Highly novel contributions are more rare, contributions with minimal novelty are more frequent.

In Big-C creativity, the creative contribution has to make a large impact on society. For example, Csikszentmihalyi studied Big-C creativity by interviewing over 90 highly eminent people (i.e., Pulitzer Price winners, professional athletes, scientists, etc.) that made large creative contributions to society. In contrast, little-c has to make some contribution to society, but the impact does not have to be large, and mini-c creativity does not have to have any impact on society at all [6]. Thus, mini-c creativity’s contribution (only to one’s own life) is a similar formulation to the P-creativity defined by Boden, in which the novelty is personal. To help differentiate little-c creativity from mini-c creativity, consider the example of photography: taking photographs that are made available to the public through a display in a local gallery or home is an example of little-c creativity. Capturing personal photographs that are solely about learning, reflection, and the creative process itself, and where the photographs are not shared publicly, would be an example of mini-c creativity. Thus, mini-c is largely about the personal, creative experience. Finally, another category

is Pro-C creativity or ‘professional expertise.’ Kaufman and Beghetto proposed that Pro-C creativity represents creative contributions from people who are highly skilled in their craft but may never achieve Big-C creativity [35]. An example of a Pro-C contribution may be an influential research paper or industrial product design. These are important to society and have reasonably high novelty but do not reach the level of extreme novelty or significant impact on society at large, but rather have impacts within their domains.

While there are creative endeavors that occur without the support of digital tools, due to technology’s increasingly pervasive role in society, many types of creative endeavors are influenced by, supported by, or completely enabled through technology. In this research, I am interested in studying users doing creative work, regardless of where their creative contributions land within the Novelty-Impact space (Figure 1). While it is clearly most beneficial to society to support creative contributions that have high novelty and high impact, it is not possible to know ahead of time which contributions will fit in that part of the space. Thus, my philosophic approach is democratic: by improving tools that are used to support Pro-C, little-c, and mini-c creativity, it is my hope to enable the emergence of Big-C creativity. However, it is important to note that it would be very difficult and risky to measure how well a CST supports creativity by waiting to see whether it enables a Big-C creative contribution.

My research approach is to measure how well tools support people engaged in mini-c, little-c, and Pro-C creativity, as these are more frequent activities. In order to improve tools that are used for all of these types of creativity, there needs to be metrics that are general enough to evaluate how well any particular tool of creativity.

2.2 Concepts Relevant to Creativity

The concepts of exploration, play, and flow are highly relevant to open-ended creative activities, especially in the context of creativity support tools. According to Boden, exploration is critical to the creative process [7]. In particular, she said that exploration within a set of rules or constraints is crucial in that it allows people to see possibilities that they might not have otherwise considered, which in turn may lead people to discover new ideas. Exploration within a set of constraints is relevant to interface constraints within CSTs, in that users are only able to explore a creative space within the boundaries of the tool itself. However, people often appropriate tools and use them in ways not intended by the developers [44]. Similarly, Boden noted that creativity and play are highly related because both are commonly open-ended creative activities, and they both are in opposition to the concept of ‘work.’ In the context of children, Amabile said that constraints are a balancing act: predictability is important but there has to be a limit to help maintain intrinsic motivation [3].

Similar to creativity, play is also an ill-defined concept, and its main characteristics are: (1) Intrinsic motivation, (2) Attention to means, rather than ends, (3) Active engagement, (4) Freedom from external rules, (5) Non-literality, and (6) Behavior dominated more by the individual than by the environment [56]. Within HCI, a play-related study by Read et al. reported dimensions of children’s fun during play, including engagement and endurability [55]. I regard these dimensions from Read et al. to also be relevant to the study of CSTs, because of their applicability to the study of play. The term ‘endurability’ is defined as the willingness to continue

or repeat an activity, and it is likely that wanting to repeat the activity reflects enjoyment of the activity. Enjoyment is likely impacted to some extent by how much the user is intrinsically motivated by the task [56, 4]. Intrinsic motivation as a creative personality facet has also been studied by Amabile [2, 4].

Csikszentmihalyi's [15] *flow* is clearly relevant to creativity [58]. Flow refers to an 'optimal experience' that happens during an activity. While 'being in flow' does not necessarily indicate that someone is creative, flow can enhance the creative process. The concept of flow is defined by 9 characteristics: (1) There are clear goals every step of the way, (2) There is immediate feedback to one's actions, (3) There is a balance between challenges and skills, (4) Action and awareness are merged, (5) Distractions are excluded from consciousness, (6) There is no worry of failure, (7) Self-consciousness disappears, (8) The sense of time becomes distorted, and (9) The activity becomes autotelic (i.e., meaning of the activity is within itself). These concepts of exploration, play, and flow, were part of the foundation of the Creativity Support Index, which will be discussed in Chapter 3.

2.3 Creativity Measurement Approaches

There is an enormous body of research on measuring creativity or various aspects of creativity, with many different methodologies and metrics for doing so. The majority of creativity studies fall into one of five approaches: psychometric, experimental, biometric, historiometric, and biographic. In addition to these approaches, some computer science and artificial intelligence researchers have begun to develop computational methods for creativity evaluation [46, 51].

Plucker noted that the psychometric approach is the most influential to the study of creativity [53]. The psychometric approach is concerned with the development of self-report, standardized surveys for measuring creativity through rigorous reliability and validity methods. In general, the psychometric approach views creativity as a quantifiable mental trait. Plucker said: “a majority of work dealing with creativity relies on psychometric methods – the direct measurement of creativity and/or its perceived correlates in individuals” [53]. He further said that the majority of creativity research employs methods that are psychometrically based. Since the psychometric approach is the most relevant to the Creativity Support Index, this approach is furthered discussed in Section 2.3.1.

Creativity is also studied using the experimental approach, in which researchers study cognitive processes of people as they engage in creative tasks. There are three primary characteristics of this approach: controlled environments, quantitative measurements (often from psychometric tools), and cognitive task analysis [53]. The Biometric approach is similar to the psychometric approach in that the goal is to quantify creativity using physiological measurements, such as electroencephalography (EEG) and galvanic skin response (GSR). In 1999, Plucker noted that there had been very little work done in this area. In fact, references that he made to this approach were actually investigating the relationship between biometrics and intelligence, rather than biometrics and creativity. More recently, I have also been doing work in this Biometric space. In Chapter 4 and Chapter 5, I will discuss my research that involves a psychological approach to evaluating creativity support tools using EEG, GSR, and other measurements [11, 12]

Another quantitative approach is the Historiometric approach. The data collected in this approach is from historical documents about highly creative people, rather than from self-reported data. These historical facts include factors such as birth order, childhood trauma, family background, etc. [53]. The Biographical approach is similar to the Historiometric approach because it also focuses on people who have demonstrated Big-C creativity; however, this approach is highly qualitative, involving interviews and archival research [31]. While these two approaches vary drastically from the psychometric approach, they are often used as sources of validity for developing self-report surveys, such as surveys that ask about past creative behaviors.

2.3.1 Psychometric Approach to Creativity Measurement

The psychometric approach to creativity measurement primarily consists of self-report techniques and rating scales for external judges [53]. Accuracy and validity of surveys are limited due to self-reporting issues, but they are beneficial because they are cost efficient, easy to administer, and can provide quantitative results. There are many types of psychometrics tools for measuring creativity, such as those that ask about divergent thinking (i.e., Torrance’s Test of Creative Thinking [61]), creative attitude measurements, and inventories that ask about past creative behavior and activities (i.e, Creative Behavior Inventory [29]). The most commonly used are tests of divergent thinking, which are especially prevalent in the education domain. Tests like divergent thinking are often used to classify how creative a person is; however, this is problematic because divergent thinking tests are not a full measure of creativity. Rather, they represent only one dimension, which means they are not suitable for

classifying people. For example, divergent thinking does not necessarily take into consideration all aspects of creativity, such as creative synthesis or artistic creativity.

External judges are also often used in creativity research to rate a person's overall creativity or to rate a person's creative product [30]. The way that judges are used varies, ranging from the types of judges used (experts versus non-experts) to the definitions of creativity provided to the judges. Most often, expert judges are used, although many studies have also utilized non-experts. In the education domain, it is common to use peer judges. External judgment studies also differ in the creativity definition that is provided to judges. Judges are sometimes provided with elaborate definitions of creativity with several dimensions to rate, but in other studies, no definition is provided at all. Karlin et al. [34] found a correlation of 0.97 between defined and undefined creativity ratings, suggesting that it does not matter if a definition is provided, as the definition did not have any effect on their results [53]. It may seem risky to not provide a definition of creativity, but Amabile argued for this approach because allowing people to use their own definition of creativity helped them to be more consistent and gave more reliable results [5, 53]. However, it is important to note that reliability does not guarantee validity in the judges' ratings.

The main criticism of external judgments is the lack of discriminant validity between other constructs. For example, judges may rate individual constructs like intelligence, competence, and creativity, but they may all be highly correlated [30]. If judges cannot discriminate one construct from another, this implies that only one construct is being measured. Hocevar was even more troubled by the results from the artistic creativity domain, where judges were unable to distinguish between technical

skills, aesthetics, and creativity, which is problematic because these constructs are not interchangeable [30]. He further said, “Since judges have trouble discriminating creativity from other attributes, it follows that they will have even more trouble discriminating various dimensions of creativity.”

Across all creativity measurements, another concern is the lack of convergent validity when researchers use multiple methods [30]. If a study were to utilize divergent thinking tests, external judges’ ratings, and creative attitude measurements, past research shows that they would most likely be loosely correlated [16, 21, 30]. This causes creative people to be ranked in different orders: a person who scores high on one test could score low on another test, implying that each test is actually measuring a different construct [30]. For this reason, Hocevar wrote that tests of divergent thinking and creative attitude measurements are actually measuring correlates of creative behavior, rather than an actual measure of a person’s overall creativity. In order to measure a person’s creativity, he argued for using inventories that ask about past creative behavior because these metrics were developed based on the activity and behaviors of highly eminent people. He also noted that the best way to predict future creative behavior is by examining past creative behavior.

Therefore, in ranking participants as to how creative they are, it is critical that researchers be selective about the measurements that they employ. If the researcher is truly studying divergent thinking, then such a test would be appropriate. However, in the case of studying artistic creativity, a measurement like the Creative Behavior Inventory [29] may be more appropriate, as it measures creativity by asking about past experiences and achievements.

CHAPTER 3: THE CREATIVITY SUPPORT INDEX (CSI)

As discussed in Chapter 1, there are no ‘convenient’ quantitative metrics available for evaluating how well creativity support tools (CSTs) actually support creativity, and furthermore, creativity itself is also very difficult to define and measure (Chapter 2). To begin to address these challenges, this chapter presents a new quantitative measurement tool, the Creativity Support Index (or CSI), which is a psychometric survey-based tool, designed to help researchers evaluate CSTs. The purpose of the CSI is to evaluate how well a tool supports creativity by measuring six dimensions that are important to creativity support tools: Results Worth Effort, Exploration, Collaboration, Immersion, Expressiveness, and Enjoyment. In Section 3.2, I will discuss how these dimensions were derived in the development of the CSI.

In productivity software evaluations, there are many quantitative metrics to employ, often relating to time and error measurements. As discussed in Chapter 1, these measures are not sufficient to evaluate creativity support. Survey metrics are also commonly used in HCI evaluation; however, until the CSI, there has been no survey metric designed specifically for evaluating CSTs, which means that researchers often borrow surveys from other domains, and these borrowed survey tools may not be appropriate and valid when applied to CST evaluation. To this end, my dissertation is focused on developing computational approaches for evaluating CSTs.

In this chapter, I present the Creativity Support Index itself and its extensive development. The CSI has been developed and validated through a rigorous psychometric process, and it is grounded in literature on creativity, play, and flow (Chapter 2). Over the course of the CSI’s development, it has been used in studies with 13 different CSTs with over 120 participants. In particular, the aspect of collaboration and how it relates to creativity support tools has been a focus. This chapter is organized as follows: I begin by presenting the final version of the Creativity Support Index, and in doing so, I discuss how to administer and score the CSI, how to report CSI results, and scenarios that demonstrate how to fit the CSI into an experimental design. To make the survey easier for researchers to administer, I also present an electronic version of the CSI. After presenting the CSI, I discuss the rigorous, iterative development of the CSI as a psychometric tool. Finally, I identify limitations of the CSI and address critical topics for researchers using the CSI, such as situations in which researchers may want to modify the CSI.

3.1 The Creativity Support Index

The Creativity Support Index is a psychometric survey that was designed to assess the ability of a digital creativity support tool (CST) to support the creative process of users by measuring dimensions that are important to supporting creativity. In this section, I present the final version of the CSI; however, I begin with motivation for the CSI’s survey structure. This section ends with a discussion around scenarios of usage for the CSI. The development of the CSI will be discussed later, in Section 3.1.2.

3.1.1 Survey Structure

In developing the CSI, I was inspired by the popularity and survey structure of the NASA Task Load Index (TLX). The TLX is a standardized survey used to quantify workload [27]. In particular, the TLX was originally developed for evaluating workload tasks in aircraft simulations and other similar human-machine equipment. The TLX has been reliably used (and even adapted) by many researchers over the past 20 years [26] and also used within a variety of domains, including the HCI community [28, 41]. Because of its widespread use, many researchers are already familiar with the tool. This means that TLX scores have meaning to researchers, and thus TLX results can be reported without excessive narrative or explanation.

The survey structure of the TLX was also appealing, since it generates both factor scores and a single, weighted score out of 100 for workload (in the TLX a higher score indicates higher workload). The TLX includes six factors: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. It is designed as a weighted metric so that it can be applied in situations where one or more of those factors may be less relevant. Similar to creativity, workload is a complex phenomenon that is understood intuitively but is not easily captured by any single, simple metric. In the TLX, there is one rating scale statement for each of the six factors. In addition to six rating scales, there is also a paired-factor comparison test, in which each factor is compared against every other factor for a total of 15 comparisons. The participant chooses which factor in each pair was most responsible for workload. These paired-factor comparisons are useful because they make the

overall index score more sensitive to the factors that are the most applicable in a given situation, since the formula for the score is weighted based on the results of the paired-factor comparison. Similarly, the CSI uses rating scales and paired-factor comparisons to generate a single, factor-weighted score out of 100, with a higher score indicating better creativity support. It is also worth noting that the paired-factor comparison information alone may be beneficial to researchers by allowing them to understand which factors are most relevant to workload, in the case of the TLX, or creativity support, in the case of the CSI.

3.1.2 The CSI: Final Version

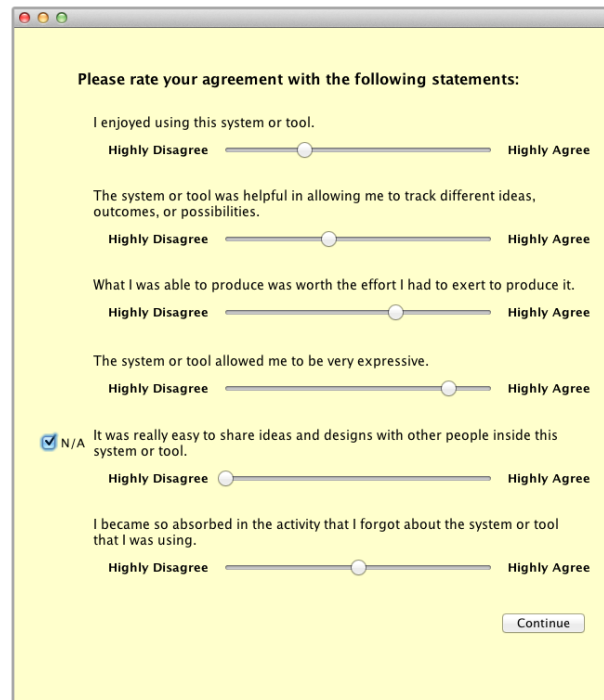
The Creativity Support Index consists of six factors: Collaboration, Enjoyment, Exploration, Expressiveness, Immersion, and Results Worth Effort. For each factor, there are two agreement statements (i.e., rating scale items). In Table 2, these agreement statements are shown. Each statement should be rated by participants on a scale of ‘Highly Disagree’(1) to ‘Highly Agree’ (10). Similar to the NASA TLX, there is a paired-factor comparison section that consists of each factor’s description being paired up against every other factor’s description 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 3. For administering the CSI, I created an electronic version: Figure 2 shows the agreement statements (i.e., rating scales) and Figure 3 shows two screenshots of the paired-factor comparisons.

Table 2: 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.

<p>Collaboration</p> <ol style="list-style-type: none"> 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.
<p>Enjoyment</p> <ol style="list-style-type: none"> 1. I would be happy to use this system or tool on a regular basis. 2. I enjoyed using the system or tool.
<p>Exploration</p> <ol style="list-style-type: none"> 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.
<p>Expressiveness</p> <ol style="list-style-type: none"> 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.
<p>Immersion</p> <ol style="list-style-type: none"> 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.
<p>Results Worth Effort</p> <ol style="list-style-type: none"> 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.

Table 3: 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



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

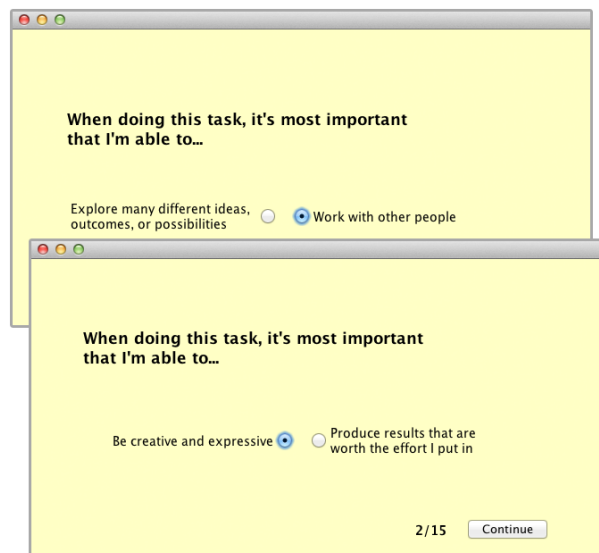
The system or tool allowed me to be very expressive.
 Highly Disagree Highly Agree

☒ 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

[Continue](#)

Figure 2: Screenshot of 6 out of 12 agreement statements in the CSI's user interface.



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 [Continue](#)

Figure 3: Screenshot of 2 out of 15 paired-factor comparisons in the CSI's user interface.

3.1.2.1 Survey Administration

When administering the CSI, participants should complete the rating scale section for each creativity support tool that they use. After rating the 12 agreement statements for each CST being studied (Table 2), participants complete the paired-factor comparisons (Table 3). In other words, the paired-factor comparison is generally only completed once and after using all relevant CSTs, as the paired comparisons are relevant to the task, rather than the tool. Of course, this setup assumes that each user is performing the same task with each tool (i.e., in a within-subjects study). If the user is performing multiple tasks with each tool being studied, then the researcher should administer the paired-factor comparison section after each tool/task combination. In Section 3.1.3, I discuss other usage scenarios of the CSI.

It is entirely possible to administer the CSI on paper; however, we recommend that researchers administer the CSI electronically. To this end, I have developed an electronic version of the CSI, which is available at my website². A benefit to using the electronic CSI is that the application will score each test automatically with results saved in a comma-separated file (csv), labeled with a participant ID and condition number (in the case of repeated measures). In the user interface, there is an initial page that allows the researcher to set up 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

²Electronic CSI available at: <http://www.erincarroll.net/csi>

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

Figure 4: Equation for manually scoring the CSI

with six items per page, as seen in Figure 2. There are also 15 pages of paired-factor comparisons with one pair per page (Figure 3). The design decision to have one factor per page was made to force participants to consider each pairing independently and to prevent participants from lining up their answers to the paired comparisons. In the TLX, the paired comparisons were routinely presented together on one page, which may have negatively impacted how participants responded.

3.1.2.2 Scoring the CSI

The electronic version of the CSI scores the surveys automatically, generating a single CSI score out of 100 for the tool being used, as well as individual factor scores that can help a researcher understand how well a tool supports various aspects of creativity. 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 with a higher score indicating better creativity support. Essentially, a CSI score is measuring the CST, the task, and the user type (i.e., user expertise with a CST). The equation for calculating a CSI score is available in Figure 4 for researchers who might wish to deploy the CSI on paper and need to score survey responses manually.

Table 4: Sample reporting of CSI results in a repeated measures study.

Factor	Avg Counts n=12	Avg. Rating, Tool X n=12	Avg. Rating, Tool Y n=12
Results Worth Effort	4	9.0	4.5
Exploration	0	4.5	5.0
Collaboration	2	4.0	4.0
Enjoyment	4	8.0	4.0
Expressiveness	1	8.5	5.0
Immersion	4	6.0	5.5
Overall CSI Score		72.33	46

3.1.2.3 Reporting CSI Results

The Creativity Support Index was developed as a measurement tool for the academic community. In order for the CSI to provide continued benefit to researchers, it is important for researchers employing the CSI to be very explicit about how they used the tool when reporting CSI results. For example, researchers should report when and how often their participants completed the paired-factor comparison test. It is also important that researchers always report the overall CSI score. Below I present an example reporting of CSI results from a fabricated, repeated measures study comparing creativity support across two different tools.

Participants generated an overall CSI score of 72.33 for tool X and an overall score of 46 for Tool Y. Table 4 shows the average factor counts for the task and average factor ratings for the two different tools. The average counts are the number of times participants chose that particular factor as important to the task, and the highest count possible for any particular factor is 5, indicating that participants chose it as more important

than every other factor. Looking at the average counts, we see that for the particular task being studied, the Results Worth Effort, Enjoyment, and Immersion factors are most important. Tool support for Exploration and Expressiveness have much less importance. Collaboration is somewhat important. Next, we can interpret the ratings. An average rating will be between 0 and 10, with a higher number indicating the tool better supports that aspect of creativity. We see that Tool X supports Results Worth Effort, Enjoyment, and Expressiveness well, but does not support Exploration or Collaboration particularly well. Tool Y has mediocre support across all factors. Overall, we see that for the given task, Tool X offers better creativity support, but could perhaps be redesigned to better support Immersion and Collaboration. Tool Y does not do a particularly good job of creativity support for this task at all, and is possibly not a good choice of tool for the task. Since a CSI score is intertwined with user expertise, we also investigate whether there is a correlation between CSI scores and users' expertise.

Clearly, a tabular format is just one choice for presenting these results. Also, in cases with large N, researchers should run statistical analysis on the data, calculating standard deviations, and statistical significance between results from different tools. In some instances, researchers may choose to manually calculate the CSI without a particular factor. In this case, it is important for the researcher to report scores both with and without that particular factor. In cases where the CSI is modified, it

is critical for researchers to explicitly discuss their modifications and to make note that they are using a modified version of the CSI, rather than the psychometrically validated version.

3.1.3 Usage Scenarios

The CSI is designed to be flexible and to apply to a wide variety of domains and tools where users are engaged in creative activities. It is also designed to be flexible in terms of how it is administered by researchers. In this section, I provide a number of scenarios to demonstrate how the CSI could be used.

- Tool Comparison, Same Creative Task, Repeated Measures: Researchers are interested in evaluating a creativity support tool by comparing it to another tool, so they designed 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 creative task in each tool. Since the creative 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 Creative Task, Between Groups: This is similar to the case above, except that participants only use one of the tools. This may be appropriate if significant learning transfer is expected from one tool to the next. In this case, participants will complete the CSI, including the pairwise factor comparisons, once after using the assigned tool. For reporting purposes, the factor counts should be reported for each group, as they may differ.

- **Multiple Creative Tasks in the Same Tool:** A CST 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 in their CST. After each creative task, a participant will complete both the agreement statement section and the paired-comparison section. Participants must complete the paired-comparison section multiple times in this scenario because the researchers are studying different tasks within a CST, and the paired-factor comparison is focused on the task. The CSI data is analyzed by looking at the difference in scores for each task. The researchers are also interested in comparing the results of the paired-factor comparisons 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 type of study, participants use the same CST for each study session and then complete the CSI at the end of the session. In this scenario, the researchers are interested in the changes in 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 creativity support tool that does not support collaboration, and subsequently, there are no collaborative tasks. Her first plan is to remove all the collaboration statements from

the CSI. However, she decides to administer the complete CSI, allowing participants to mark the collaboration statements as non-applicable. 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 could be beneficial to participants. In other words, while the tool does not support collaboration, participants selecting collaboration frequently in the paired factor comparisons would indicate that participants wished that the tool would support collaboration.

- **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 post-experiment survey. In this case, the researcher calculates the CSI and reports it as a comparison metric for the use of other researchers that are studying similar CSTs.
- **Individual Rating Scales:** While the CSI provides an overall index of creativity support, researchers are comparing two different CSTs and are mostly interested in differences between the factor scales for 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.

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 ceilings, 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							
Intrinsic motivation				✓			
Attention to means rather than ends				✓			
Active engagement				✓			
Freedom from external rules	✓						✓
Nonliterality							✓
Behavior dominated more by person than environment			✓		✓		

Figure 5: The relationships between the primary creativity theories and related concepts, which resulted in the six factors on the Beta CSI.

3.2 Development of the Creativity Support Index

3.2.1 Beta Version

The development of the CSI began by creating a beta version that was exclusively based on the literature on creativity and creativity support tools, as discussed in Chapter 2. Since many of the concepts presented overlapped, it made sense to synthesize them (Figure 5). Through a card sorting process, I grouped these theoretical concepts into six factors that were used for the Beta CSI: Exploration, Collaboration, Engagement, Effort/Reward Tradeoff, Tool Transparency, and Expressiveness. I then wrote one agreement statement per factor for a total of six items (Figure 6). In contrast to the final CSI version presented in Section 3.1, the beta CSI was much shorter. The electronic version also had a different user interface: the final version of the CSI allows users to mark collaboration items as non-applicable but this option was not available in the beta CSI. All of these changes will be explained throughout the documentation of the CSI's development process.

The beta CSI was used as a mechanism for gathering early feedback about the survey, and it was deployed in three user studies. The goal of these deployments was to pilot test the beta version to see if people understood it and felt it was appropriate for studying creative tasks and creativity support in systems or tools. Overall, I found that the beta CSI was easy for participants to complete, but two of the factors, Collaboration and Tool Transparency, were confusing to some participants.

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

Figure 6: 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).

3.2.1.1 Study 1: Ken Burns Study

The beta CSI was first deployed in a within-subjects experiment in which 32 people created photographic slideshows using the Ken Burns Effect. This effect allows people to select regions of interest in a photograph, and software is used to animate the image by interpolating the viewpoint between the selected regions of interest, allowing a narrative video to emerge. I used the beta CSI to compare two different interaction techniques for specifying Ken Burns regions. One of the techniques 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 panning and zooming and was based on the Ken Burns Effect interface in Apple’s iPhoto 2008. After completing a series of trials with each technique, participants completed both the CSI and the NASA TLX.

Since this study did not involve collaboration and the tools provided no support for collaboration, I 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 9 out of 32 participants selected the lowest values. In addition, a few participants verbally asked if they should ignore the item. Based on these results, it seemed probable 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*. It is also possible that while participants did not work with other people on this task, they may have imagined that it was possible to work with others.

In addition to the CSI, participants also completed the NASA TLX so that I could do a cross comparison on the appropriateness of the survey to the task. Participants were asked to indicate whether they felt that the CSI or the TLX was the most appropriate evaluation survey for the slideshow creation activity. The CSI was reported as the most appropriate survey for the task by 21 participants, while only 10 participants said that the TLX was the most appropriate. I also asked participants to indicate which survey (i.e., CSI, TLX, both, or neither) was confusing to them. This question was asked because the TLX may be confusing to people after completing a creative task, since the TLX was designed for assessing tasks with clearly defined

goals. Fourteen participants said the TLX was confusing to them; four participants said both the TLX and the CSI were confusing; and 13 participants said that neither were confusing.

3.2.1.2 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. This study was primarily a qualitative user study in which participants spent one hour using a bimanual color exploration tool that was embedded in a drawing program, and they were told to think aloud as they used it. At the end of the study, participants were instructed to complete a paper version of the beta CSI.

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 that the question was referring to a tool within the interface actually disappearing from view inside the application.

3.2.1.3 Study 3: Kinematic Templates Study

Lastly, 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 interesting kine-

matic drawing effects [20]. Since this was a longitudinal study, each participant was involved in four or five sessions over a course of three to twelve weeks, with each session lasting approximately one hour. After each session, participants were given a paper version of the CSI to complete.

Similar to Study 1 and 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.

Since this was a longitudinal study, I also looked at the CSI scores for participants over time. Figure 7 shows that there was a general trend in the overall CSI scores increasing over time, indicating that as participants developed more expertise with the tool, it was better able to support their creative work. This is a positive outcome, as it makes sense for a tool to better support creativity as users gain expertise.

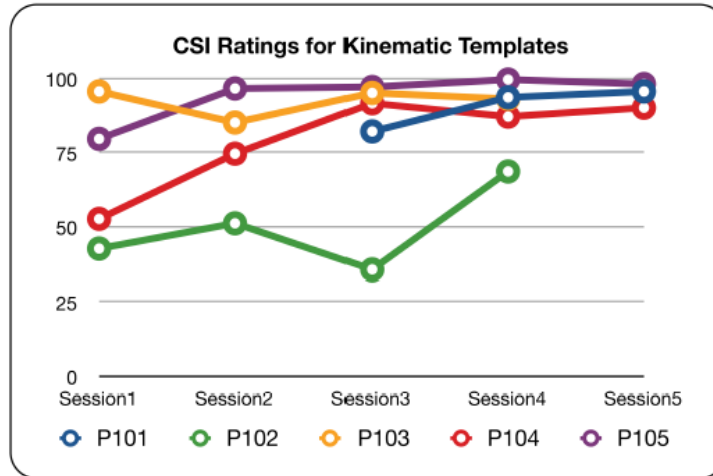


Figure 7: Scores on the beta CSI from a longitudinal study with Kinematic Templates. In general, these scores increased over time.

3.2.2 Creativity Word Study

Based on the overall positive response to the beta CSI, I proceeded to further develop the CSI in a formal psychometric process to produce a statistically reliable and valid instrument. The experiments with the beta CSI helped to identify several issues that needed to be addressed. First, people may use different terminologies than those used in the creativity literature and in the beta CSI. Second, I wanted to be more thorough in the creativity factor categorization than the card sorting process may have allowed. Therefore, in an effort to improve the CSI, I followed 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 asked them how much each word was related to workload [27]. After that, the TLX developers analyzed the data using a principle components analysis, and they renamed each extracted component to represent a preliminary TLX factor that was representative of the words perceived to be related to workload.

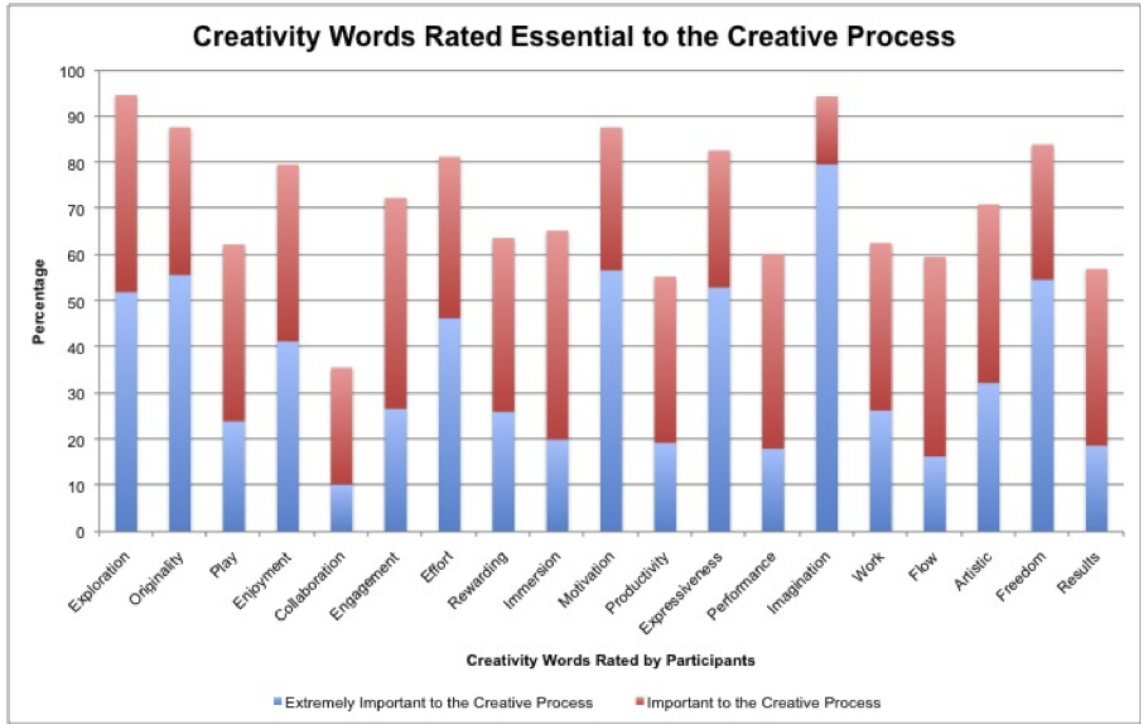


Figure 8: Participants (n=300) rated how much each of these 19 words were related to the creative process. The purpose of this study was to refine the factors used on the CSI.

3.2.2.1 Methodology

I conducted a creativity word study and factor analysis, similar to a study in the NASA TLX development [27]. In this study, 300 people 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.’ Participants were recruited using Amazon’s Mechanical Turk ³, which allows people to complete short tasks for a small monetary incentive. In this study, participants were paid ten cents each to complete these ratings. The 19 words that I selected were both from creativity research and from common parlance. These 19 words and ratings are available in Figure 8.

³Amazon’s Mechanical Turk: <http://www.mturk.com>

Following a process similar to Hart and Staveland [27], I conducted a principle components analysis using an extraction method based on components that had an Eigenvalue above 1.0 (i.e. the Kaiser rule) [33]. Using this method, there were six extracted components, which coincidentally was the same number of factors used in the beta CSI.

3.2.2.2 Results

I named the six extracted components based on the words that loaded the most strongly for that component. These named components became the six factors for the final CSI and are available in Table 5. I excluded the components that had ‘motivation’ and ‘imagination’ as the strong loading words because the CSI is a metric that focuses on the *tool*, rather than the *person*. While the CSI is not intended to measure intrinsic motivation, it is likely that intrinsic motivation will have an impact on individual’s CSI scores, in particular around Enjoyment and possibly Exploration. Therefore, researchers may also want to employ surveys that measure facets of creative personalities. This is a clear example of why it is essential to use mixed-methodologies when studying creativity support tools. Instruments for measuring the creative personality were discussed in Chapter 2.

In Figure 8, 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. As a result, collaboration did not load strongly on any component. It is likely that collaboration did not reflect highly in these ratings because a very strong stereotype exists that the creative genius works

alone [31]. However, even though collaboration was not rated particularly important, I strongly agree with the creativity research on the importance of collaboration [31, 57]; thus collaboration remains as a factor on the CSI.

The results of this study led to the components (or scale names) that are now used in the final version of the CSI: Results Worth Effort, Expressiveness, Exploration, Immersion, Enjoyment, and Collaboration. Table 5 specifies how these new factors are related to the original names in the beta CSI.

3.2.3 Lengthening the CSI

An initial goal of the CSI was for it to be similar in form and length to the NASA TLX [27]. However, based on best practices in psychometric research, I later realized that it was imperative to have multiple statements for each factor. Without this structure, it would be impossible to consider a factor as its own scale. There needs to be at least two items per factor in order to compute the reliability of each scale (although generally, reliability actually increases by having even more items in a scale). Being able to calculate the reliability of a scale is important not only when developing a psychometric tool, but it is also important to researchers that want to calculate their own reliability ratings, as relevant to their user study.

To address reliability issues, I lengthened the CSI to have two agreement statements (or items) per factor. While even more items per factor could have further increased the reliability, I wanted to be cautious about the length of the CSI, as it is important for research participants to be able to complete the CSI quickly, especially in the case of repeated measures studies.

Table 5: 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 principle 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

3.2.3.1 Methodology and Results

In order to lengthen the CSI, I created an extended version by writing additional rating 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 this case, I 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 3.2.2. Since the purpose of the study was to find the best rating scale statements, the paired-factor comparisons were not relevant, so only the rating scale statements were deployed. Similar to the beta CSI, all items were rated on a scale of ‘Highly Disagree’ (1) to ‘Highly Agree’ (10).

This temporary, extended version of the CSI was deployed 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 survey was deployed in a classroom or workshop setting, where the participants had tasks to complete. Therefore, participants were not given any specific tasks; instead, they were allowed to continue working on the activities that they were presently engaged in. They were instructed to fill out the CSI after they were finished using their creativity support tool. The advantage of this method is that I was able to collect data across a variety of tasks and tools,

which makes these results more generalizable. It is important to note that none of these CSTs were collaborative, although it is possible that some students may have collaborated on their tasks outside of the tool.

I used a factor analysis (FA) to analyze the data, which is similar to a principle components analysis but is the preferred method in psychometrics for understanding the relationship between survey items. In the FA, the Kaiser rule was used again for the extraction method with an oblique rotation because there was no reason to believe that the factors were not correlated. I began by running FA's on survey items for each of the six scales. For example, I performed an FA on the seven items that were written for the Results Worth Effort scale; an FA on the eight items for the Exploration scale; etc. The purpose of running an FA on the items in each individual scale was to verify that the items in each scale only represented one factor. Some items loaded strongly (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 was intended; therefore, these items were discarded.

After using a factor analysis to verify one factor per scale, I 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, I 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, I was able to set higher standards because only the two best performing questions for each scale were needed. After reducing each scale using this criteria, I 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 two items per scale that I was left with are shown in Table 6, along with reliability results.

3.3 Test-Retest Reliability

A test-retest study was conducted to test the stability and reliability of the CSI over time, using the two-item factor version presented in Table 6. All rating scale statements were rated on a scale of ‘Highly Disagree’ to ‘Highly Agree.’ 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, I expected to see slight changes in scores over time, unless using experts as participants or using very simplistic creativity support tools.

3.3.1 Methodology

Twelve participants with an interest in sketching were recruited to participate in this test-retest study. In recruitment, all participants were aware that the study involved returning three weeks later to repeat the same experiment. I selected 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, I expected that learning on this tool would be 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 then were asked to spend 30 minutes sketching anything that they liked. Participants were allowed to choose what they wanted to sketch, as

Table 6: The CSI was lengthened to have two items per factor. In this table, I 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

Table 7: Average, standard deviations, and test-retest reliability from the overall CSI score and each of the 6 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

assigning them something to draw that they were not interested in may have impaired their ability to be creative. After sketching, each participant completed an electronic version of the lengthened CSI (i.e. two statements per factor) from Figure 6, and the paired factor comparison. All participants came back three 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. Participants were paid \$5 for the first session, and \$10 for the second session.

3.3.2 Results

Overall, the test-retest reliability results were very good. The reliability of the overall CSI score from Session 1 to Session 2 was 0.65. I 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. These reliability results are available in Table 7.

3.3.3 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. Intuitively, some participants expressed that they wanted to mark it as ‘Not Applicable,’ but this was not an option. Instead, many people 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 three weeks went by, many participants answered this question differently than in their first session.

The results of this study indicated that there were issues with the Collaboration scale and that more work was needed in this area. Specifically, I realized that we needed to: (1) Re-visit the items in the Collaboration scale; (2) Allow for ‘N/A’ responses in the CSI’s user interface; and (3) Test the CSI in a collaborative tool. These improvements to the Collaboration scale will be discussed later in Section 3.5.

3.4 User Studies with the CSI

In addition to the test-retest study, the CSI from Table 6 (i.e. CSI with two-items per scale) was incorporated in several user studies. In this section, these user studies are described, and in Table 8, I have summarized the results and reliability information for each user study. For reliability, I computed the Cronbach’s Alpha for

each scale as a measure of the internal consistency of the scale. I have also reported the descriptive statistics for the overall CSI score and each of the six scales. For the overall CSI, the average score is out of 100 with a higher number indicating better creativity support. For each of the six factors, the score is an average based on the sum of the two statements for each scale with a maximum score out of 20.

3.4.1 Study 1: Adobe Photoshop

I visited an adult education class on Adobe Photoshop that was taught at The Light Factory (a local museum of photography in Charlotte, NC) and had students complete the CSI in response to the activity that they were working on for this course. There were five female participants between the ages of 25 and 55. This took place on the last day of class and students completed the CSI at the very end, after they were finished with their final project in Photoshop. The average CSI score for Adobe Photoshop in this study was 84.20 (SD=18.84).

3.4.2 Study 2: AutoDesk Sketchbook Express

The CSI was also used as an additional measure in my work that investigated the temporal, creative work process, using self-report and physiological measures [12]. In this study, the task was open-ended sketching with AutoDesk Sketchbook Express while wearing electrodermal activity (EDA) sensors and an electroencephalography (EEG) headset. There were 11 student participants who said that they enjoyed sketching. Seven of these participants were Fine Arts students or in a design-related degree. The average CSI score for SketchBook Express was 64.79 (SD=17.06). This study is discussed in more depth in Chapter 4.

Table 8: Descriptive statistics and Cronbach's Alphas for all six CSI scales for Study 1, Study 2, and Study 3 in Section 3.4. Also included are the averages and standard deviations for all three overall CSI scores. *Alpha

	Study 1		Study 2		Study 3	
	Avg (SD)	*	Avg (SD)	*	Avg (SD)	*
Results Worth Effort	17.20 (4.21)	0.97	14.45 (4.61)	0.83	15.95 (3.24)	0.87
Exploration	15.80 (3.27)	0.47	12.27 (4.08)	0.52	15.40 (3.75)	0.79
Enjoyment	17.20 (4.15)	0.95	12.27 (3.85)	0.85	15.10 (4.32)	0.79
Expressiveness	17.40 (3.71)	0.87	14.82 (3.74)	0.84	15.90 (3.63)	0.86
Immersion	15.80 (5.36)	0.92	10.73 (4.52)	0.57	14.90 (3.86)	0.75
Collaboration	16.40 (4.16)	0.97	9.36 (3.08)	0.28	9.85 (1.98)	0.80
Overall CSI Score	84.20 (18.84)	n/a	64.79 (17.06)	n/a	76.52 (16.25)	n/a

3.4.3 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 [22]. 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 was 76.52 (SD=16.25).

Several participants in this study indicated some confusion about the Collaboration statements. They asked questions such as, “But what if the question is not applicable?” This result was not surprising, given the issues previously found with the Collaboration scale.

3.5 Collaboration

After the test-retest study in Section 3.3 and the CSI user studies in Section 3.4, it was clear that improvements needed to be made to the Collaboration scale for several reasons. First, I had yet to test the CSI in a study that actually involved a collaborative activity and a tool that explicitly supported creativity. This is partly due to the fact that tools which support collaboration are much less prevalent than single user desktop tools. The fact that the CSI was not tested with a collaborative activity and/or tool most likely contributes to the low test-retest reliability for Collaboration, as reported in Section 3.3. This led to a second concern: since the CSI had yet to be tested on a collaborative activity and a CST supporting collaboration, it was possible that the Collaboration items selected from the item-level analysis in Section 3.2.3 were not truly representative of the best questions for the Collaboration scale. Therefore,

in an effort to improve the Collaboration scale of the CSI, I conducted a final user study with the goal of finding the best Collaboration items by testing the CSI on a collaborative activity and tool.

I also made improvements to the CSI's user interface. While improving the scale and testing on a collaborative CST is critical, it still does not solve how participants should respond to statements about collaboration, when collaboration is not applicable. Therefore, I made interface improvements to the electronic version of the CSI, which allowed participants to select an 'N/A' check box when appropriate.

In this section, I will discuss the collaboration study, from which I derived a new Collaboration scale, as well as interface improvements for handling 'N/A' responses.

3.5.1 Collaboration Study

3.5.1.1 Methodology

A collaborative, creative writing study was conducted using Google Docs. There were 16 participants in this study, who were recruited through Amazon's Mechanical Turk. Each participant collaboratively wrote a creative story with a 'confederate participant,' rather than another study participant. In other words, our Turk participants were led to believe that they were writing a creative story with another Turker. However, as a control variable, participants were paired with a confederate instead. The confederate in this study was a student, Jordan Bullington, pursuing a Master of Arts degree in creative writing.

When a person visited the study posting on Mechanical Turk, they were told that they would be collaboratively writing a creative story on Google Docs and that they

would complete a short exit survey (i.e., the CSI) when they were finished writing. They were also told that it would take no more than five minutes to be matched up with a partner from Mechanical Turk. After accepting the task, the participant was given a link to the study's Google Doc, where the confederate participant was always waiting. Therefore, this study incorporated some deception, in that participants were under the impression that they would be collaboratively writing with another study participant. This deception was important to the study, as I did not want to impose additional social pressure on the participants by telling them that they were writing with a member of the research team. Participants were paid \$5.00 for participating in this study, and as a way to motivate them and increase the "creative stakes" of participants, they were told that highly creative stories had the potential to earn an extra \$5.00. In reality, all participants received the bonus payment.

Inside the Google Doc, the participants were instructed to collaboratively write for 30 minutes, in response to a creative writing prompt. The writing prompt included two different photographs⁴ that were unique and did not seem to belong together (Figure 9). 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 30 minutes of 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 version that participants completed in this study was modified from the two-item per scale version in Table 6. For collaboration, participants were actually

⁴Photographer: Christen Lesley Lucas, <http://www.sassyfrassstudios.com>



Figure 9: 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. Photo Source: Christen Lesley Lucas³

given a total of four agreement statements. These additional collaboration items came from the extended version of the CSI that was used to perform the item-level analysis in Section 3.2.3. Specifically, from that item-level analysis, I took the four collaboration items that performed the best (i.e., highest corrected item-total correlation). This allowed me to essentially perform the item-level analysis on the Collaboration scale again.

3.5.1.2 Analysis and Results

In order to finalize the Collaboration scale, I performed an item-level analysis on the four collaboration items used in this study. Similar to the item-level analysis in Section 3.2.3, I first dropped the question with the lowest CITC and then dropped the other question by looking at item content. Specifically, since only one of the statements was a negative question (i.e. reversed rating scale), I decided to drop that item. In this study, our Collaboration scale had a Cronbach's Alpha of 0.914, and the selected statements are available in Table 9.

Table 9: The collaborative, creative writing study on Mechanical Turk resulted in these agreement statements for the Collaboration scale.

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.

3.5.2 User Interface Improvements

Throughout the majority of the 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 I first began the CSI development, I expected participants to mark ‘Highly Disagree’ for statements in which collaboration was not involved, which would automatically score Collaboration statements as a zero on a scale of 0-20. However, participants consistently said that they wanted to mark this item as non-applicable. Therefore, in the final version of the CSI (as previously shown in Figure 2), I added a check box beside the collaboration items, which allows participants to mark these items as non-applicable. When the ‘N/A’ check mark is ticked, the slider automatically re-positions itself at ‘Highly Disagree,’ thus coding Collaboration statements as zero.

3.6 Discussion

3.6.1 Mixed Methods Evaluation

The Creativity Support Index is a post-experiment survey metric that I developed and validated for the purpose of quantifying creativity support. Despite its iterative and rigorous development process, I do not expect the CSI to be the only metric

used by researchers. In other words, researchers should have a variety of research methodologies at their disposal, and the CSI is intended to be one additional research tool within an ontology of methodologies for evaluating CSTs.

3.6.2 Modifying the CSI

There are some circumstances in which researchers may want to modify the CSI. However, in order for the CSI to be developed as a standardized, psychometric tool, it went through a rigorous development process. If a researcher was to modify the CSI, it may lessen the reliability and validity of the CSI. Before modifying the CSI, I recommend that researchers reflect on these modification scenarios.

3.6.2.1 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 their CST, and this modification is perfectly acceptable.
- Removing Collaboration: Many researchers will be studying CSTs that do not incorporate collaboration. As previously discussed in Section 3.1.3 (i.e. Usage Scenarios), I recommend that researchers do not remove the Collaboration statements. The CSI has been designed to handle 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 com-

parisons, in order to keep one standardized version of the CSI. Standardization of the CSI is very important. First, by having a standardized survey that fits both collaborative and non-collaborative tools, it allows researchers to easily compare CSI scores on CSTs with collaboration and CSTs without collaboration. By having a standardized survey, it also makes it easier to interpret research articles that use the CSI, because the overall index score will have a greater meaning, if there is only one version. It is also the case that the paired-comparison section can be beneficial to non-collaboration 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 the tool does not currently support collaboration. For all of these reasons, it is my recommendation that the Collaboration scale is not removed from the CSI.

- **Removing Any Scale:** Similar to Collaboration, I do not recommend that researchers remove any scales on the CSI. The main purpose of the paired-factor comparison section 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 overall 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. In this case, the researcher should report why a particular scale was removed and how the modified score was calculated, in addition to reporting 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.
- **Re-Wording Questions:** It is strongly recommended that researchers do not re-word any questions on the CSI. Not only will researchers not be able to report the CSI's overall index score (as in the situation above), but researchers will also not be able to claim that they are using a standardized rating scale from the 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. In this case, 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 a modified version of the CSI has been employed.

3.6.3 CSI in the Creative-Arts

The Creativity Support Index has been developed and tested in a research environment that is strongly influenced by multidisciplinary research projects within a

variety of digital arts. While I have tried to ensure that the CSI is general and is not only relevant to creativity support within the arts, the CSI has not been extensively tested in other creative domains, such as engineering design or computer programming. However, I do not consider this a major limitation, since the factors in the CSI (Expressiveness, Exploration, Enjoyment, Results Worth Effort, Collaboration, and Immersion) should be relevant to these other domains. What might be needed are additional tools that look at some other aspects of those domains not captured by the CSI, which is why I reiterate that the CSI should be one of multiple tools used for creativity support evaluation.

3.7 Future Work

There are many future studies that can be conducted in order to understand the boundaries of the Creativity Support Index. However, these suggested studies would make contributions that extend beyond investigating the contributions and limitations of the CSI. This future research would enable CST researchers to further refine their understanding of methodologies and theories in creativity support tools.

- Studies that investigate to what extent that CSI scores are a function of a person, a function of the creative task, a function of expertise, or a function of the tool itself.
- Studies that investigate whether CSI scores vary as expected, by either altering the capabilities of a tool (i.e. removing a functionality from the tool), or comparing a known ‘good tool’ to a known ‘bad tool.’

- Studies that compare CSI scores to results from a traditional usability evaluation, such as heuristic evaluation, to see how much the scores correlate. In other words, what does the CSI provide to researchers that existing techniques may not?

CHAPTER 4: IN-THE-MOMENT-CREATIVITY (ITMC): STUDY 1

A primary challenge in evaluating creativity support tools is that, in general, we are not able to detect *when* a person is being creative. Some creativity research focuses on the relationship of physiological measures to creativity, but much of this previous work has taken the stand that creativity is being measured as long as the experimental task is a ‘creative task’ or if the participant has been classified as a creative person [48]. These assumptions lead to results in which physiological signals are labeled as indicative of creativity throughout a capture session, just because the person or task is considered creative. However, it seems evident that in the course of working on a creative activity, any user will experience periods of high creativity and some periods of low creativity; and these periods could exist for both people classified as highly creative and people who have not been classified as creative.

The goal is to approach a level of granularity that allows researchers to study the temporal, creative work experience, or ‘in-the-moment-creativity’ (ITMC). If it is possible to reliably detect when a person is experiencing creativity, then we could:

1. Research which CST features lead users into or out of creative experience peaks,
2. Develop adaptive interfaces that try to keep people in a creative experience peak for a longer time, and
3. Quantify environmental or contextual impacts on the creative experience during CST usage.

Thus, successfully measuring ITMC has the potential for high pay off in the future. However, ITMC measurement is a challenging area due to issues of validity. In other words, are we actually measuring a construct related to creativity? It is my position that ITMC does not equal creativity but rather represents a correlate of creativity: that of the personal, temporal experience of feeling highly creative during a work process. This is the correlate of creativity that is the most relevant to CST evaluation, as the goal is to design CSTs that promote and sustain periods of high personal, creative experience.

In this chapter, I present a computational approach to measuring ITMC, which involves triangulating several temporal metrics, including self-report ratings of creativity, external judgments of creativity, and physiological measurements, as captured by electroencephalography (EEG).

4.1 Background: Physiological Metrics for Creativity

The connection between bodily reactions and emotional experiences was first discovered by William James [32]. The most accepted relationship is between autonomic nervous system arousal and emotional reaction: emotion is represented by both arousal (or activation) and valence, in which arousal is typically measured using electrodermal activity (EDA) [40]. EDA measures skin conductance through the eccrine sweat glands, often using sensors worn on the fingers or wrist. Research in affective computing [52] and HCI [42, 47] has shown interesting correlations between EDA and self-reported ratings of arousal and engagement. There is also a significant amount of work studying the physiological manifestations of creativity, particularly with re-

spect to brain physiology. The position of this research is that there is a relationship between creativity and cortical arousal [48]. Cortical arousal can be measured using amplitude from the alpha-wave power band through electroencephalography (EEG). In other words, high alpha-wave amplitude represents low cortical arousal.

In Martindale's review of physiological creativity research, he reported several relationships between cortical arousal and creativity [48]. The main trends include: (1) highly creative people have higher overall cortical arousal; (2) highly creative people have more variability in cortical arousal; and (3) highly creative people exhibit lower cortical arousal during highly creative moments, compared to less creative people. The latter trend is believed to occur because highly creative people are able to defocus their attention, whereas non-creative people may focus their attention more and thus spike their arousal. While Martindale reported that these trends are found across many studies, he also noted that a large majority of studies are not able to find statistical significance for these trends, while a few other studies yield entirely opposite results.

From Martindale's work, it is evident that there is some relationship between creativity and cortical arousal, but the inconsistencies and lack of statistical significance are concerning. It seems highly possible that the relationship between creativity and cortical arousal is confounded across several variables: (1) the 'creative activity' being studied; (2) the method of classifying people's creativity level; (3) the stage of creative activity being measured; and (4) participant priming. In the first study by Martindale and Hasenfus, the creative activity was a creative writing assignment in response to a writing prompt, given to students in a creative writing class [48]. In

this study, creative people were classified as highly creative or less creative based on how two course instructors rated the creativity of the writing assignment. They found that highly creative people actually had *lower* baseline arousal (although there was no significant difference), which goes against the trend that creative people have higher overall cortical arousal. This study also found that highly creative people had significantly lower cortical arousal than less creative people during creative inspiration but not during creative elaboration; so this result was highly dependent on the stage of the creative process.

Martindale and Hasenpus's second study involved an activity where people engaged in different types of speech: random speech where they would say random words that came to mind and fantasy speech where people would tell fantasy stories in response to a prompt [49]. Random speech represented the inspirational phase, whereas fantasy speech represented the elaboration phase. Half of the participants in this study were told to be as creative and original as possible, but the other half were not directed in any way. Creativity levels of participants were classified into four groups based on the results of the Alternate Uses Test, Remote Associations Test, and Shipley Institute of Living Vocabulary Test. This study found that highly creative people had lower cortical arousal during creative inspiration; *however*, this was only found in the group where people were instructed to be creative. This suggests that creative priming impacts behavior in these types of studies.

The other concern in comparing creativity studies like these, and such comparisons are the basis for trends reported by Martindale [48], is the difference in the creative activity being studied and how creativity is measured. For example, it is unlikely

that researchers would measure the same facet of creativity in a study about drawing or a study about creative writing or a study about problem solving. Furthermore, given the complexity of measuring creativity itself, as discussed in Chapter 2, it is concerning to compare studies where creativity is being classified in different ways, given that each form of creativity measurement is more than likely measuring something different or a different correlate of creative behavior altogether, as hypothesized by Hocevar [30]. This suggests that researchers need to be clear in identifying the type of creativity correlate that they are measuring.

4.2 ITMC Measurement Approach

Measurement of ‘in-the-moment-creativity’ during a creative work process has the potential for high pay off in both the design and evaluation of creativity support tools. However, given the complexities and challenges in creativity measurement, physiological creativity research, and the lack of an exact definition of what constitutes creativity, ITMC measurement and interpretation must be approached cautiously. Therefore, ITMC measurement should be viewed as a method for detecting a *correlate* of creativity based on user experience, rather than a quantified metric of creativity itself. This differentiation does not lessen the importance of assessing the validity of ITMC ratings. However, it is important that there is some level of assurance that a time segment identified by a participant as being highly creative, was in fact a period in which something specific happened that was experienced as creative by the user. In other words, if the participants has a similar experience later, would they also label it as highly creative? Since the goal of using ITMC measurement is to

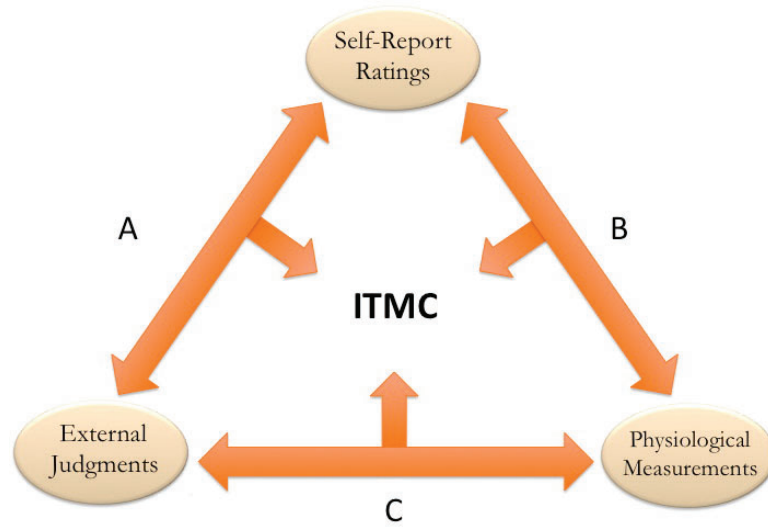


Figure 10: The first approach to measuring ITMC involved a triangulation of three temporal metrics: self-report ratings of creativity, external judgments of creativity, and physiological measurements (e.g., EEG)

study how to make CSTs better support the creative process, a consistent correlate is needed.

In this first attempt at measuring ITMC, multiple sources of measurement were used: physiological measures, self-reported periods of creativity, and externally judged periods of creativity. These sources were expected to correlate in such a way that there was relatively high reliability for ITMC ratings. Thus, as shown in Figure 10, relationships were expected between self-reported ITMC and external judges' ratings of ITMC (Relationship A); between the physiological measurements and self-reported ITMC (Relationship B); and between physiological measurements and externally judged ratings of ITMC (Relationship C).

4.3 ITMC Study 1

I conducted an experiment to investigate whether it was possible to measure 'in-the-moment-creativity' (ITMC) with a triangulation approach. The creative activity

in this experiment involved sketching with a digital tool. Sketching was selected because there are many drawing tools available that do not require a high degree of specialization, and sketching is a reasonably popular pursuit, making it feasible to recruit participants. A highly specialized CST would introduce potential concern, in that it would cause a large variation in skill levels between participants, unless experts were used. It is also important to note that while sketching is an open-ended, artistic activity that is likely to lead to some periods of high creativity, this is not always guaranteed.

The first data source for capturing ITMC was electroencephalography (EEG), which is a measure of the electrical activity (or voltage) from the surface area of the brain (i.e., a reading from the scalp). The other source was electrodermal activity (EDA), or a measure of eccrine sweat gland activity. Participants wore both EEG and EDA while sketching and being screen recorded; however, there were technical errors with the EDA hardware so only EEG was utilized in the analysis. After sketching for 30 minutes, participants watched their screen recording video and retrospectively indicated when they were being creative using a custom video application that I developed. These self-reported ratings of creativity were the second triangulation sources for identifying ITMC. External judges also watched the same sketching videos from each participant in order to identify and rate segments when the participants were being creative. These external judgments were the third triangulation source for identifying ITMC. In the following section, I present the materials and methodology involved in this user study.

4.3.1 Materials

4.3.1.1 Sketching Tool

Participants (n=11) sketched using AutoDesk’s SketchBook Express with a Wacom graphics tablet. This drawing program was selected because it has a simple interface that offers a variety of brushes and features (e.g., smudge, blur, etc.), and it was relatively easy to learn. None of the participants had used this particular tool before, but all participants seemed to learn the application very quickly. Two participants had never used a graphics tablet before.

4.3.1.2 ITMC Reporting Application

In order to temporally self-report creativity, participants used a custom video player application that allowed them to retrospectively identify time periods experienced as creative. The motivation for having participants self-report retrospectively was because concurrent self-reporting would distract from the creative activity. It was my expectation that having participants watch screen capture videos of themselves sketching (with audio included) would allow them to recall what they were thinking and feeling as they sketched, similar to a retrospective think-aloud study [24].

While watching a video of their sketching activity, participants indicated that they were being creative using a keystroke. In Figure 11, the ‘Start’ label indicates that a participant has pressed a key to indicate that a creative moment has begun. When the participant felt that their creativity was changing, they would push the same keystroke to indicate the end of the segment, which would cause a red rectangular region to

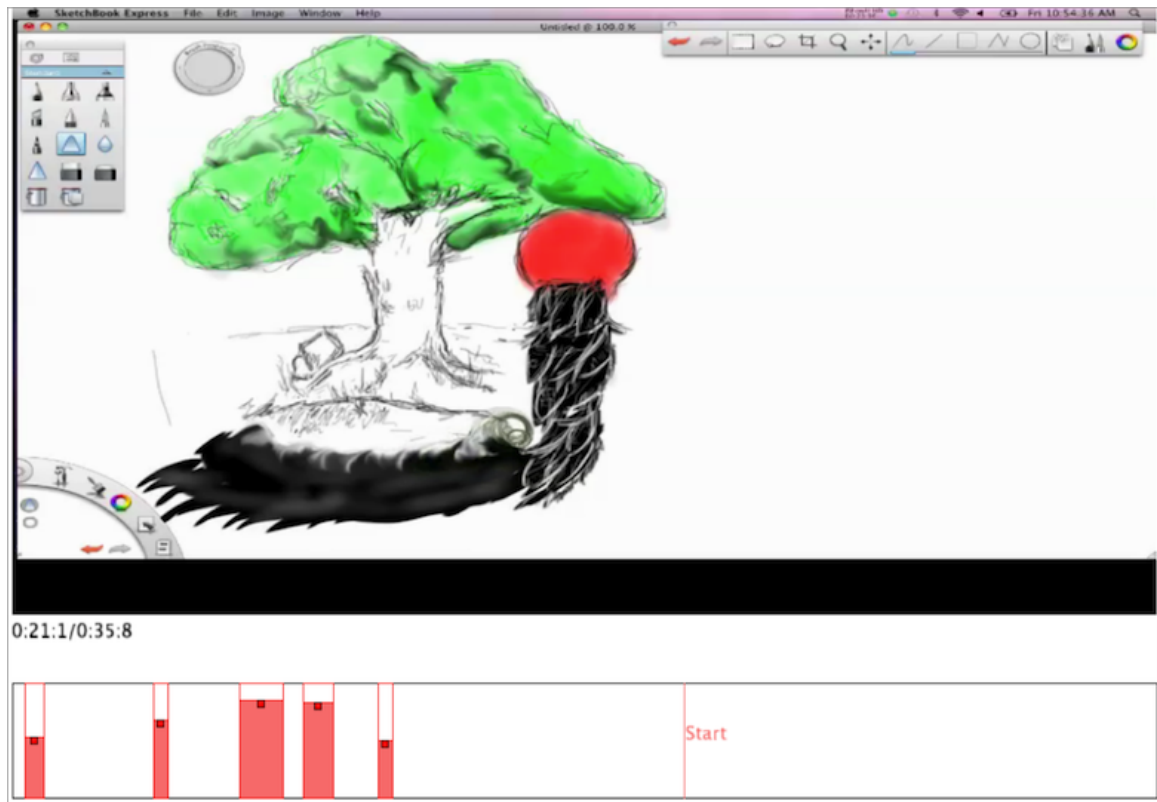


Figure 11: Participants rated their creativity by watching their screen recordings inside this custom application. They used keystrokes to indicate when a period of creativity begins and ends, which is represented by each rectangle. Afterwards, they adjusted the creativity level using the red circle inside each pink rectangle.

appear. After each creative time period was marked, the participant adjusted the pink circle inside the rating rectangle to indicate how creative they were being. The default position of this circle was centered vertically in the rectangle, representing moderate creativity. They were briefed on the meaning of this circle and were shown how to use it to indicate their level of creativity, by dragging it up and down. The creativity level in each creative time period was mapped from 1-100. Time segments in the sketching activity that were not identified as a creative experience were coded as zero.

4.3.1.3 Emotiv EEG Headset

The Emotiv headset is a 14-electrode, wireless EEG headset⁵, designed for the brain-computer interface community and promoted heavily towards gaming. At each electrode position, EEG provides a measure of the electrical activity of that surface area. EEG was selected because it is a sophisticated measure that provides more information than a linear measure like EDA. EEG has also been commonly used in many of the physiological creativity studies. The Emotiv headset, in particular, was selected because it is an off-the-shelf option with far less setup time than a traditional, medical-grade device. Traditional EEG may require up to two hours of setup time and the use of electrode gel, whereas the Emotiv headset takes about 15 minutes to setup and uses saline solution instead. Figure 12 shows the Emotiv headset.

Due to the complexity of EEG data, the data analysis involved machine learning, which will be discussed in Section 4.5.2, and is a common approach to handling EEG data. For example, Grimes et al. used machine learning on EEG data when studying working memory load [23]. The main motivation for using a machine learning technique is because there is not enough consistent information in the EEG/creativity research to allow us to know exactly what to look for in an EEG signal to identify ITMC. I used self-reported ratings of creative experience and external judgments of creative experience, as labels (or ground truth) in the classifiers and then tested how accurately those classifiers performed. With this goal in mind, many features were extracted from the EEG data in order to have enough information for classification.

⁵Emotiv: <http://www.emotiv.com>



Figure 12: The Emotiv EEG headset (left) and Affectiva Q sensor (right) were used in the ITMC study to capture physiological responses. The Emotiv headset is a wireless EEG device, and the Q includes EDA, skin temperature, and a 3-axis accelerometer.

Thus, the approach to EEG signal analysis was predominantly a ‘black box’ approach, rather than being based on functional neuroscience. The only exception to this rule is that I extracted alpha and beta bands because research shows relationship between activity on these bands to creativity and mental effort [23, 48].

It is common to filter out artifacts (e.g., eye blinks, facial movements, etc.) prior to extracting EEG data, but I chose to not remove artifacts. Ekman’s research shows a relationship between facial expressions and emotion [19]. Similarly, Mikhail et al. [50] used ‘noisy’ EEG data because filtering out data from facial movements could remove some information that corresponded to that emotion. This approach makes sense to evaluating CSTs, as I do not want to filter out any positive or negative emotions, as this information may be important for classification. It is important to note, however, that using noisy EEG data can make classification more challenging.

In pre-processing the EEG data, a 0.16 Hz high-pass IIR filter was applied to each participant’s EEG data file in order to remove the DC offset from the signal. Then,

Fast Fourier Transforms (FFTs) were applied to every two seconds of data using a Hamming window function. The EEG headset’s sampling rate is 128 Hz, therefore FFTs were run on every 256 samples. In each two-second epoch, I rejected any epochs that contained any data point in the sample that exceeded a fixed threshold. According to Delorme et al., this is the most accepted and effective method for detecting excessive eye blinks or eye movement artifacts [17]. Each participant’s session was approximately 30 minutes, which means approximately 900 FFTs were run for each person. Within each FFT, features were extracted from the alpha band (7.5-12.5 Hz), low beta (12.5-25 Hz), and high beta (25-30 Hz). The extracted features used for each FFT across all bands include: measure of power across 14 electrodes; and measure of asymmetry, cross-spectral densities, and coherence for each of the 7 symmetric pairs. Asymmetry measures the relative amount of activity in the left versus right hemisphere of the brain; cross-spectral density measures the relationship between different brain regions; and coherence measures the linear dependency between different brain regions.

4.3.1.4 Affectiva Q Sensor

Participants also wore the Affectiva Q on their dominant hand [54]. This device has three measurements: EDA, skin temperature, and a 3-axis accelerometer. The benefit of using the Q sensor is its form factor. It is worn on the wrist like a watch, which makes it very comfortable for the participant to wear and requires no setup time. The sensor is shown in Figure 12 on the right side. Unfortunately, the Q model used in this study was a prototype, and it shut off during three sessions. Therefore,

the Q's data has been excluded from this analysis, but it was used as a measurement in a later study described in Chapter 5

4.3.1.5 Creative Behavior Inventory (CBI)

The CBI is a 77-item inventory for assessing past creative behaviors [29]. The items on this inventory give specific examples of creative activities, ranging from arts and crafts to science and math, and respondents are asked to indicate how many times they have performed each activity in their adult life. I used the CBI as a way to categorize how creative the sketching participants were. To score the CBI, I added up the responses across all 77 questions and then normalized the scores to a scale of 0 to 100. Participants completed the CBI after setting up the EDA and EEG hardware but before they began sketching.

4.3.1.6 Creativity Support Index (CSI)

After sketching with SketchBook Express, participants completed the CSI (Chapter 3, CSI Version from Figure 6). The CSI was used as a way to evaluate how well SketchBook Express supported the creativity of the participants. The CSI produces an index score out of 100, where a higher score indicates better creativity support.

4.3.1.7 Post Experiment Survey

Participants also completed a post-experiment survey at the very end of the session. The questions asked, along with their rating scale, were:

1. Please rate how creative you felt overall while sketching during this study.

Not Creative (1) to Extremely Creative (10)

2. Please rate how creative you typically feel when sketching.

Not Creative (1) to Extremely Creative (10)

3. Please rate how comfortable you felt when asked to self-report your creativity.

Not Comfortable (1) to Extremely Comfortable (10)

4. Please rate how accurate you felt when asked to self-report your creativity.

Not Accurate (1) to Extremely Accurate (10)

5. Please rate how easy or difficult it was to use the video interface for reporting your creativity.

Not Easy To Use (1) to Extremely Easy To Use (10)

4.3.1.8 Post Judgment Survey

Similar to the Post Experiment Survey, this survey was completed by all judges after rating each video. These questions have the same scales as the survey above.

1. Please rate how creative you felt the participant was being overall during this video.
2. Please rate how comfortable you felt rating the creative process of this participant.
3. Please rate how accurate you felt rating the creative process of this participant.
4. Please rate how easy or difficult it was to use the video interface rating the participant's creativity.

4.3.2 Procedure

4.3.2.1 Sketching Participants

I recruited participants from the University of North Carolina at Charlotte, who said that they enjoyed sketching. Participants were from a variety of disciplines, as I did not want to limit this study to exclusively art students. There were a total of 11 participants, and seven of these participants were art students or in a design-related degree. One person also said that sketching was their creative outlet, and another person said that sketching was part of their job. Participants were paid \$15 for the 90 minute study.

When participants first arrived, the EEG headset and the Q sensor were set up. The Q was placed on each participant's wrist, but the EEG headset was more complicated since each of the 14 electrodes had to be wet with saline and then aligned properly on each participant's head. It took approximately 15 minutes to setup the EEG for each participant. Once the equipment was setup, participants completed the CBI, which took about 5-7 minutes. It was during this time that baseline physiological measures were captured. After this, participants were given a tutorial of SketchBook Express and then given 30 minutes to sketch; this was screen recorded and audio was also recorded. Participants were allowed to sketch anything that they wanted, and they were told this in advance of the study in case they wanted to come prepared with ideas. I allowed them to sketch anything that they wanted in order to avoid impairing their creativity by assigning something that they were not interested in.

After participants were finished sketching, they took the CSI to assess SketchBook Express and then they rated their creativity by watching the screen recording video of their sketching activity inside a custom video application. Before proceeding to self-report, I gave them a brief tutorial on how to use the application, which involved showing them how to indicate creative time segments and also how to rate their creativity level using a visual rating scale inside each time segment (Figure 11). They were verbally told that the rating scale controlled by the red circle (inside each pink rectangle) could be moved between *Low Creativity* and *High Creativity*.

As participants watched their video inside the self-report ITMC application, they were instructed to push the spacebar when they felt that they were being creative in the video and to push spacebar again when they felt that the creative segment had ended. Then, they were instructed to adjust how creative they felt during this segment using the visual rating scale. As previously discussed, I did not provide participants with a definition of creative experience, as allowing participants to use their own definition increases reliability [5]. This rating process was then repeated multiple times throughout the video playback in order to capture and rate all creative moments. They were not instructed on how often they should indicate creative time segments.

Finally, participants filled out the Post Experiment Survey, which was previously discussed. This survey was given in order to assess how creative the participant felt overall and how accurate and comfortable they felt when self-reporting their creativity.

4.3.2.2 External Judges' ITMC Ratings

Three external judges were recruited through the recommendation of professors in the Department of Art and Art History at the University of North Carolina at Charlotte. Two judges were advanced seniors in their applied digital art degree program, and the third judge had already completed her degree in that area. All were practicing artists. Judges were paid \$12 per hour to rate 11 videos, which lasted 30 minutes each.

Before beginning the project, the external judges were told that they would be watching videos of people sketching and that they would be temporally rating the creative process of those people. Similar to Amabile [5], external judges were not given any definition of creativity. However, they were told to rate people's creative process, rather than rate a creativity judgment of the end product. Given the focus on everyday creativity and the idea that personal creativity can occur at all skill levels, judges were told that they should try to ignore the skill level of the participant because some participants may have more technical drawing abilities than others. All three judges watched each of the 11 participants' sketching videos, during which they judged creativity using the same ITMC reporting application. This made the process of rating ITMC very different for judges in comparison to participants, since the judges rated the videos based solely on the actions that they observed in the videos, not knowing what was going on inside the participants' minds as they sketched. As a result, I expected this difference between judges and sketching participants would lead to lower correlations in Relationship A (Figure 10). After watching each video,

they also took a Post Judgment Survey, which was similar to the Post Experiment Survey that the sketching participants completed.

In analyzing the reliability of external judges, I calculated the consistency between judges (or inter-judge consistency) using a two-way mixed Intraclass Correlation Coefficient (ICC). ICC analysis produces two correlations: the reliability of using one single judge, if you were to choose one of the judges in the set (ICC Single), or the reliability of the average ratings of all judges (ICC Average). Both the ICC Single and the ICC Average are useful in different circumstances. In general, when we are interested in how reliable the raters were as a whole, the ICC Average is the relevant measure. In contrast, when we want to know the reliability of using just one of the judges, the ICC Single is the relevant measure. In this case, I wanted to examine the correlation between the average of all judges' ratings and the self-report creativity ratings, so I first calculated the ICC Average.

4.4 Hypotheses

For the first hypothesis, I expected to find inter-judge consistency between all three of the judges' ratings of creativity ($H1$), as measured by ICC Average, which is the reliability of using the average of all three judges' ratings. I also expected to find a correlation in the reliability between the average creativity ratings of all judges and the participants' self-reported creativity ratings ($H2$, Relationship A in Figure 10).

- $H_1 : ICCAverage_{J1*J2*J3} > 0$
- $H_2 : R_{JudgesAvgITMC*SelfReportedITMC} > 0$

The next hypotheses involve the relationships between the physiological measures and the reported creativity ratings from participants and judges, in which the goal was to classify periods of high creative experience in the EEG data from these ratings. I expected to find a pattern in the extracted EEG features between periods of low creative experience and periods of high creative experience. The discrete variables of low and high creative experience were respectively derived from the bottom 1/3 and upper 2/3 of creative experience ratings, both from self-reported creativity ratings and the average of all judges' creativity ratings. Therefore, two classifiers were created for each person, one based on the self-reported creativity ratings and the other on the judges' average creativity ratings, and I expected to find acceptable classifier accuracy for both of them. H_3 measures the classification accuracy of using EEG features with respect to self-reported creativity (Relationship B in Figure 10). H_4 measures the classification accuracy with respect to the average creativity ratings of the external judges (Relationship C in Figure 10).

- $H_3 : Accuracy_{EEGFeatures*SelfReportedITMC} > 50\%$
- $H_4 : Accuracy_{EEGFeatures*JudgesAvg} > 50\%$

4.5 Results

4.5.1 Inter-Judge Consistency

The inter-judge consistency was computed for the external judges by computing a 2-way mixed ICC. I was interested in the ICC Average, since this is the reliability of using the average of the judges' creativity ratings. The ICC Average was 0.59 ($p < 0.001$), indicating moderate reliability between judges, allowing H_1 to be

accepted. It is also worth noting that the average correlation for ICC Single, which would be the reliability of using ratings from only one judge, was 0.32 ($p < 0.001$). This indicates that it would be possible to use the ratings from just one of the judges but with a lower level of reliability.

4.5.2 Judges' ITMC Ratings and Self-Report ITMC

In order to assess the relationship between ITMC ratings of external judges and ITMC ratings of sketching participants, I computed a Pearson's correlation between the average creativity ratings of all judges and the self-reported creativity ratings of each participant. The correlation between judges' ratings and self-report was 0.25 ($p < 0.001$). While this is a weak correlation, $H2$ can still be accepted. This means that the personal, temporal creative experience self-reported by participants has some relation to the temporal, creative experiences identified by the external judges. The weakness of this relationship should not be surprising: after all, the judges' segments are likely related to creative execution, since they could not know what was going on in the participants' minds. The participants' self-reported segments likely reflect times of creative execution, as well as creative ideation.

4.5.3 Classifying ITMC with EEG Features

The investigation of physiological measurements and ITMC ratings allowed me to further assess the reliability of the ITMC ratings. The approach to testing $H3$ and $H4$ was to create a machine learning classifier for each participant using WEKA [25]. I used 10-fold cross validation to train the data and then classified using Sequential Minimal Optimization (SMO), which is a common support vector machine algorithm.

For each participant, the data was normalized to values between 0-100, and two different classifiers were created for each participant: one classifying based on self-reported ITMC ratings and the other classifying based on the average of the judges' ITMC ratings. For both types of ITMC ratings, I created a binary variable, using the bottom 1/3 of the ratings to form Low Creative Experience Periods (which includes 0-coded moments) and the upper 2/3 of the ratings to form High Creative Experience Periods.

The overall classification accuracy was 72.54% (SD=9.83) for self-reported ITMC ratings, and it was even higher at 82.98% (SD=8.62) for classifying based on the average judges' ITMC ratings. These classification accuracies were very promising, especially given that I used 'noisy' EEG data (not filtering out facial movements). These high classification accuracies for ITMC ratings allowed both $H3$ and $H4$ to be accepted. Furthermore, it makes sense that the judges' average ratings classified higher, given the judges' ratings had higher ICC reliabilities ($H1$) compared to self-report ITMC ratings from a single person ($H2$).

4.5.4 Cortical Arousal and Creativity Level

The previously discussed physiological creativity research [48] identified many trends between a person's overall creativity level and cortical arousal. Unfortunately, this is a thorny area that is filled with many possible confounds: the most important confound being how a person's creativity level is measured. The relationship between cortical arousal and a person's overall creativity is not particularly relevant to measuring ITMC, since ITMC is a temporal measure. However, it is important to

Table 10: Descriptive statistics for questions on the Post Experiment and Post Judgment Survey, completed by both participants and judges. These ratings are on a scale of 1-10.

Post Experiment & Judgment Questions	Participants' Avg Rating (SD)	Judges' Avg Rating (SD)
How creative when sketching in experiment	6.82 (1.83)	N/A
How creative when sketching typically	7.64 (1.63)	N/A
How comfortable reporting ITMC	8.45 (1.29)	8.54 (1.53)
How accurate reporting ITMC	7.13 (2.64)	8.22 (1.93)
How easy it was to use the ITMC reporting app	8.00 (1.90)	7.54 (2.44)

note that even with a small sample size, I was able to corroborate one of the trends from the research literature. Specifically, I found that during High Creativity Periods, the highly creative participants exhibited significantly higher alpha-wave power, or lower cortical arousal, ($M=23.89$, $SD=5.60$), compared to the less creative participants ($M=16.61$, $SD=2.43$), $t(9)=2.51$, $p<.05$. High Creativity Periods were defined as the upper 2/3 of ITMC ratings. Highly creative participants were those who scored 50 or higher on the Creative Behavior Inventory.

4.5.5 Survey Results

For Sketchbook Express, the average score on the Creativity Support Index (CSI) was 64.79 ($SD = 17.05$) out of 100, which is a measure of how well the tool supported creativity. This number may be beneficial in the future when comparing similar tools.

Analysis of the Post-Experiment surveys showed that both participants and judges reported high ratings for how comfortable and accurate they felt reporting periods of creative experience. Table 10 summarizes the results from the Post Experiment and Post Judgment surveys.

4.6 Discussion

The primary goal of this work was to detect *when* a person is experiencing high creativity. My approach to capturing the temporal, creative work process, or ‘in-the-moment-creativity’ (ITMC), involved a triangulation approach of three temporal metrics, including self-report ratings of creative experience, external judgments of creative experience, and EEG measurements.

The results show that by using a data triangulation approach, it is possible to identify periods of high creative experience for an individual. Using EEG data, I was able to find ITMC ratings with high classification accuracies: 82.98% for external judges’ ratings and 72.54% for self-reported ratings. Reliability of 0.59 was found between all three judges’ ITMC ratings, and reliability of 0.25 was also found between the external judges’ ITMC ratings and the sketching participants’ ITMC ratings.

The methodological setup in this work was fairly complex. Participants wore physiological sensors while sketching, which involved the initial expense of the equipment and the effort of setting up the equipment for each participant. In addition, physiological data requires more complex analysis and expertise on the part of the researcher. In terms of self-report, participants had to re-watch their entire sketching activity in order to self-report their creativity. The use of three external judges was costly and time consuming because they had to be paid to watch all 11 sketching videos. Clearly, this triangulation approach is not always ideal, and it may be more reasonable to use only one of the relationships in the triangulation approach – Relationship A, B, or C in Figure 10 – taking a small loss in reliability.

It might also be interesting to consider what can be learned based on using only one of the three relationships in Figure 10. Relationship A may be appealing because it incorporates the most traditional measures: self-report ratings and external judgments. In our study, ‘A’ performed the poorest with a correlation of 0.25. This correlation may be accepted in some cases, but this weak correlation may not be worth the cost associated with using external judges. The benefit to this approach is that physiological sensors are not required. The relationship that generated the strongest result was ‘C,’ which was between external judges and participants’ physiological measures. I used a machine learning classifier for ‘C’ and was able to classify ITMC periods with high accuracy at 82.98%. This approach was the most successful, but it was also expensive because external judges had to be paid and EEG also had to be setup for each participant. Therefore, in this case, I consider ‘B’ with a classification rate of 72.54%, to be ideal. There is a small loss in accuracy compared to ‘C,’ but external judges are not necessary. Thus, EEG with self-reported ITMC offers the best results for the least expense and effort.

Of course, it would be ideal if we were able to use just one of the three measures. Participants reported being very comfortable self-reporting ITMC, and given that these periods classified well and correlated weakly with external judgments of ITMC, it seems that the sketching participants were successful in reporting their personal creative experiences. This means that self-report alone may be a useful measure in evaluating and analyzing CST usage, and it is certainly worthy of future research. However, I do not believe that relying on external judgments of ITMC alone is a good approach to the overall goal of evaluating and analyzing CST usage, given the time

and expense incurred, as well as the fact that judges are not able to consider what is going on in the participants' minds as they were completing a creative activity with a CST.

Another option is the use of physiological measurements alone. At this time, physiological measurements cannot be used without some way to build a classifier. Some possibilities that I explore in Chapter 5, is the reliability of a person's ITMC classifier in two different sessions of a user study (i.e. repeated measures), whether an ITMC classifier works across users, and also whether an ITMC classifiers works across two different tasks (i.e. sketching and writing).

4.6.1 Potential Confounds

There are a number of confounding issues that could be at play in this data, given the complexity around creativity research:

- **Ideation vs. Execution:** This study measured 'in-the-moment-creativity' during the execution (elaboration) phase of a creative activity. Since people knew in advance that they would be able to draw anything, it is possible that the study did not capture the ideation (i.e. inspiration) phase, in some cases.
- **Creativity Stakes:** It is unclear whether the participants in this study were strongly motivated because there was no particular pressure to perform. Some may have had higher internal performing pressure than others.
- **Activity Level:** The raters may have considered the artists to be creative when they were actively sketching. It is entirely possible that creativity and activity are highly correlated, and it may be that activity level is a good proxy for study-

ing ‘in-the-moment-creativity.’ I hoped to investigate activity level using the Q sensor’s accelerometer data but that was not possible, given the technological issues with this device.

- Task Constraints: If participants were given a creative drawing ‘prompt’ with creative constraints, rather than allowing participants to sketch anything that they wanted, it is possible that participants might have been more creative. Working under constraints sometimes promotes creativity [5, 9].

CHAPTER 5: EXTENDING ITMC MEASUREMENT: STUDY 2

As a follow-up to ITMC Study 1 in Chapter 4, this chapter further investigates the measurement of ‘in-the-moment-creativity’ (ITMC) by examining ITMC measurement in two different tasks (drawing and writing) with creative activity prompts: drawing in response to caption prompts and writing in response to cartoon prompts. Five participants also came back and repeated the same study again for a re-test.

5.1 ITMC Application Improvements

One contribution of this chapter is improvements to the ITMC Reporting application. In this section, I discuss observations and results from ITMC Study 1 that helped me identify issues with the original ITMC Reporting application. Then, I discuss and present the revised ITMC 2.0 Reporting application.

5.1.1 Design Reflections and Observations

Based on observations and results of ITMC Study 1, I realized that there were three design problems with the original ITMC Reporting application (Figure 11, Ch. 4). First, I observed that participants often skipped ahead in their screen recording videos when watching, identifying, and rating creative moments. Thus, they were not watching every moment of their video. This was problematic because by not watching the entire video, participants may have missed some of their creative moments. Second, there were also a few participants who marked creative time segments that all stacked

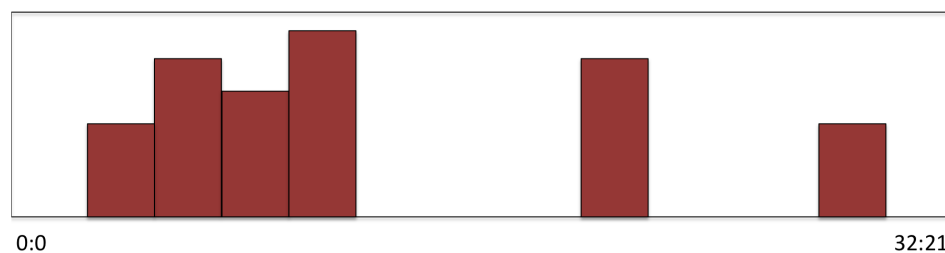


Figure 13: A mocked up example of how some participants chunked their creative rating segments in the original ITMC Reporting application from ITMC Study 1.

up against each other. Rather than being one large creative moment with a fixed creativity level, they created smaller chunks that all lined up against each other, with different creativity levels. A mock example of this is available in Figure 13. From observing participants that stacked chunks up against each other, in order to show varying levels of creativity, it seemed very tedious for them to vertically line up these chunks. This indicated that continuous self-reporting may be better for participants to show varying levels of creativity, rather than a chunking approach.

Finally, I was concerned with the amount of data from ITMC Study 1 that was coded as ‘non-creative.’ For example, in Figure 11 from Chapter 4, the data that was identified by the participant as being creative, was rated on a scale of 1-100. However, the moments that were not selected as creative, were coded as zero. The self-reporting modality of finding and selecting creative time segments resulted in an abundance of zero coded data. Since participants were completing a creative task, they were most likely already ‘primed’ to be creative, so ultimately, there were probably very few ‘true’ zero moments. Additionally, since participants did not likely watch every single moment of the video (using the chunking method), they probably failed to watch some moments that were coded as zero.

5.1.2 ITMC 2.0 Reporting Application

The original ITMC Reporting application relied upon participants identifying and rating ‘chunks’ of creative moments, which resulted in several problems, as discussed above. To mitigate these issues, I created the ITMC 2.0 Reporting application, which uses continuous self-reporting (rather than chunking) and also addresses the issue of participants scrubbing forward in the videos. A continuous self-reporting design was primarily selected based on my previous success in using Phidget sliders to collect temporal engagement ratings of audience members in performing arts [42].

To use the ITMC 2.0 Reporting application, participants watch their screen recording videos using this application. They provide self-reported ratings of creative experience using the Griffin PowerMate 3D Mouse (Figure 14). This 3D mouse, which actually resembles a knob, controls a slider widget for rating creativity on a scale of ‘Not At All Creative’ to ‘Extremely Creative’ (Figure 15). Similar to a volume knob, if the 3D mouse is turned to the right, the creativity rating increases, and if it is turned to the left, the creativity rating decreases. The 3D mouse is always locked to the slider widget in the UI. As participants rate their creative experience using the 3D mouse, the creativity ratings are drawn below their screen recording in a timeline, as seen in Figure 15.

When participants watch their screen recording videos inside ITMC 2.0, they are not allowed to forward the video, but they are allowed to rewind. To rewind the video, participants simply click on the rating timeline (where the self-reported ratings are drawn), which will then show an additional menu for playing and editing video



Figure 14: Participants in the ITMC Study 2 used the Griffin PowerMate 3D Mouse to self-report their ratings in the ITMC 2.0 Reporting application.

that has already been rated, as shown in Figure 16. Specifically, this menu allows participants to choose ‘Play Segment’ or ‘Edit Segment.’ Playing the segment will show 30 seconds of a participant’s video. Selecting to edit the segment will play the same 30-second window but will replace the ratings from this segment with new ratings that participants provide with the Griffin 3D Mouse. After playing a segment or editing a segment, a new option in the menu will appear, called ‘Next Segment.’ This will allow participants to move down the timeline in 30-second windows, playing or editing segments. Participants can select ‘Done’ to return to their traditional mode of self-reporting their creative experience, and rating will resume where participants last left off.

5.2 ITMC Study 2 Methodology

Similar to ITMC Study 1 in Chapter 4, I conducted an additional user study to investigate the detection of high creative experience. This study is primarily a follow-up to ITMC Study 1, in that I am investigating whether high creative experience can be predicted using machine learning. In many ways, the purpose of Study 2 is to

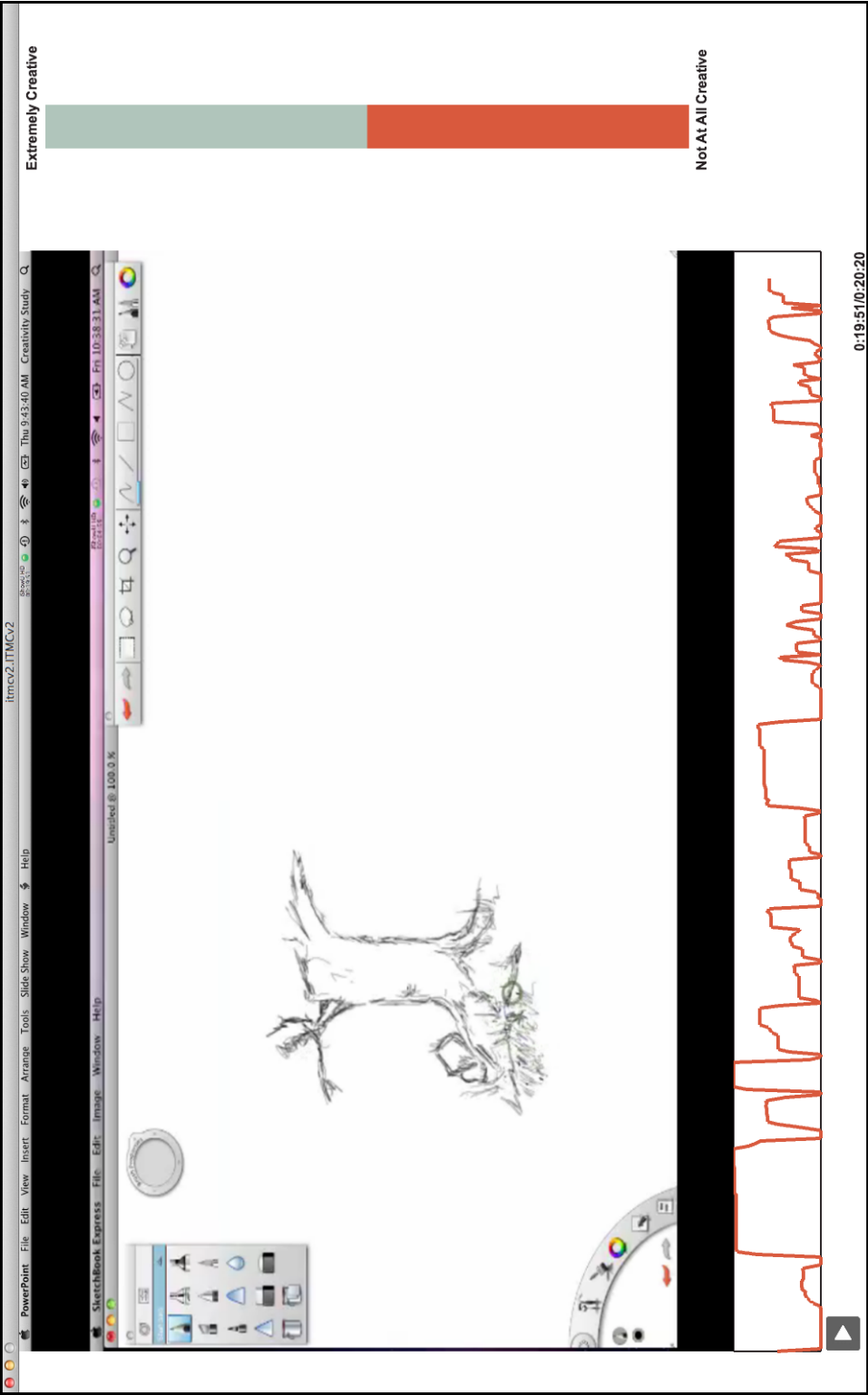


Figure 15: The ITMC 2.0 Reporting application relies upon continuous self-reporting from participants, in which the creativity slider widget on the right is controlled by a Griffin PowerMate 3D Mouse.

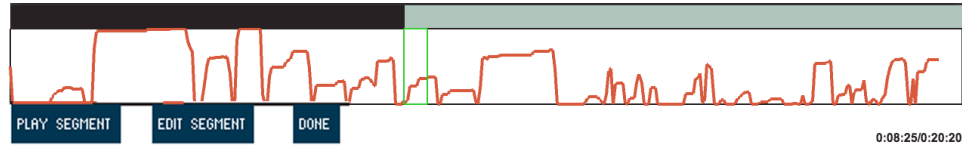


Figure 16: To rewind, participants click on the timeline, which will highlight a 30-second video segment. They can choose to ‘Play Segment’ or to ‘Edit Segment.’ After playing or editing, they can choose ‘Next Segment’ (not shown) to move to the next 30-second window, or ‘Done’ to return to traditional video rating.

investigate whether my results from Study 1 can be replicated. In contrast to Study 1, this study differs in that multiple creative activities were studied (sketching and writing), additional temporal measures were employed (EEG, EDA, Accelerometer, Keyboard/Mouse Logger), the participant sample size was increased ($n=24$ from $n=11$), and five random participants were invited back for a retest. A comparison of ITMC Study 1 and Study 2 is summarized in Table 11.

5.2.1 Materials

5.2.1.1 ITMC 2.0 Reporting Application

Similar to the ITMC Study 1 in Chapter 4, participants used a custom video application that allowed them to retrospectively rate their creative experience. In this case, they used the ITMC 2.0 Reporting application (Figure 15).

5.2.1.2 Physiological Measurements

Participants also wore the same sensors from ITMC Study 1, which included the Emotiv EEG headset and the Affectiva Q sensor (for measuring EDA and a 3-axis accelerometer). These sensors are shown in Figure 12 (Chapter 4). In contrast to Study 1, there were no technical difficulties with the Affectiva Q sensor.

Table 11: A comparison of ITMC Study 1 and Study 2. *Affectiva Q for capturing EDA was worn during Study 1 but was not used due to technical difficulties.

	Study 1	Study 2
Participants	n=11	n=24 5 re-test participants
Activity	Open-ended sketching	Sketching in response to captions Writing in response to pictures
Duration	30 min.	15 min. sketching, 15 min. writing (Multiple caption/picture prompts)
Tools	Sketchbook Express	Sketching: Sketchbook Express Writing: PowerPoint
Temporal Measures	EEG, EDA*	EEG, EDA, Accelerometer, Keyboard/Mouse Logger
External Judges	Yes	No

5.2.1.3 Activity Metrics

As discussed in Chapter 4, after ITMC Study 1, I wanted to study whether activity metrics could be a proxy for creative experience. In other words, were participants reporting that they were highly creative when there was lots of on-screen activity? This was not possible to look at in Study 1, since activity metrics were not captured. However, in ITMC Study 2, I was able to utilize the accelerometer data from the Q Sensors, in addition to a logger that I created to record the timestamp of every single keystroke and mouse movement, as another measurement of activity.

5.2.1.4 Captions and Images

I selected 20 images for participants to write captions for, and 10 captions for participants to draw pictures for. These images and captions were selected from sources that primarily included the New Yorker's Caption Contest.

Participants were not required to go through all 20 images or 10 captions. Rather, they were to spend approximately 15 minutes on sketching in response to captions, and 15 minutes on writing in response to images.

5.2.1.5 Survey Metrics

Similar to ITMC Study 1, participants filled out the Creative Behavior Inventory, as described in Chapter 4. Participants also completed a post-experiment survey that asked about comfort and accuracy in self-reporting creativity, how creative they typically feel when sketching and writing, and how easy it was to use the ITMC 2.0 Reporting application. These questions and results are presented later, in Section 5.5.6.

5.2.2 Procedure

There were 24 participants in this study. Similar to ITMC Study 1, I recruited participants from the University of North Carolina at Charlotte with an interest in sketching and writing. Participants were told in advance that they would be drawing in response to captions and writing captions in response to cartoons. Therefore, I largely relied upon self-selection bias for participants that enjoy and feel comfortable with these activities. Participants were paid \$15 for participating in this study.

When participants began the study, they first read and signed the consent form and filled out the Creative Behavior Inventory (CBI). Then I set them up with the sensing equipment: the Affectiva Q and the Emotiv EEG headset. It took approximately 15 minutes to setup the EEG headset on each participant. In my past experience using the Emotiv EEG headset, I had difficulty setting up the headset on certain hair

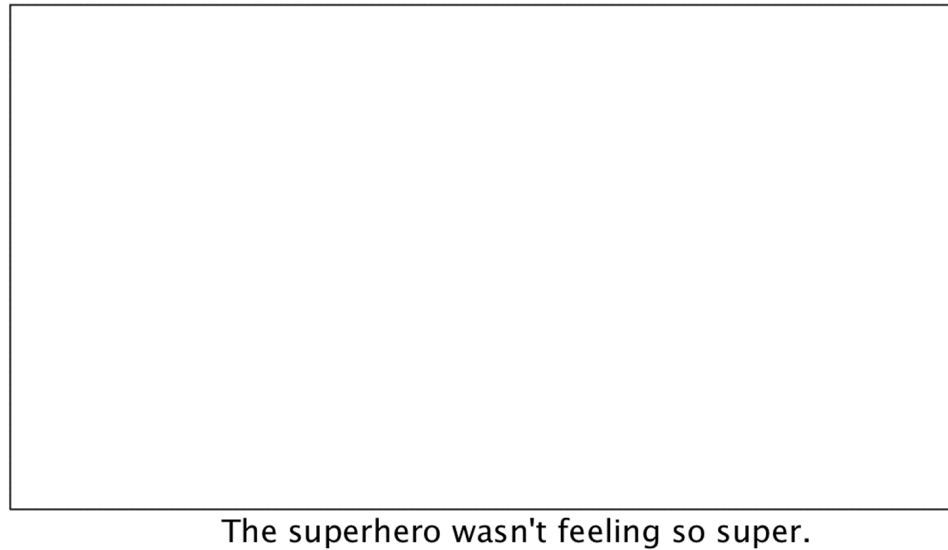


Figure 17: Example of a caption that participants sketched in response to.

styles. For example, the headset does not work with participants who have dreadlocks and afros; and occasionally, very curly, thick hair is also problematic. Therefore, in recruitment, I was more specific about what hair styles were not allowed.

The first activity was to sketch in response to caption prompts. Specifically, there were 20 caption files that opened inside of Autodesk Sketchbook Express. Figure 17 is an example of a caption prompt opened in Sketchbook. Participants were not required to sketch in response to all captions. Rather, they were given a total of 15 minutes to spend on the caption prompts (i.e., sketching tasks). After three minutes on one caption prompt, I encouraged them to move onto the next caption prompt. All caption prompts were completed in the same order.

After drawing in response to captions, participants wrote captions in response to images, inside of PowerPoint. Figure 18 shows an example of this. Similar to the sketching activity, participants were not required to write captions in response to all

20 images. Rather, they spent approximately 15 minutes on the caption activity, and after two minutes on one image, they were encouraged to continue to the next image prompt. All image prompts were also completed in the same order.

Both the sketching activity and the drawing activity were screen recorded. After completing both activities, participants used the ITMC 2.0 Reporting application to self-report their creative experience. Before self-reporting, I gave participants a demo on how to use ITMC 2.0 using an example video. Similar to ITMC Study 1, participants were not given a definition of what it meant to be creative, rather they were allowed to use their own definition of creativity, as allowing participants to do so increases reliability [5]. Finally, the user study ended with participants taking the Post Experiment Survey. Two to three weeks later, five random participants were invited to repeat the same user study. In their re-test session, a new set of images and cartoons was used.

5.3 Data Processing

This section describes how all of my sensor data was processed. Essentially, I extracted features from all sensors for every two seconds of data. This resulted in one data file (or matrix) for every participant.

5.3.1 EEG

I handled the EEG data very similarly to ITMC Study 1, in that I did not filter out facial movement artifacts. In pre-processing the EEG data, I applied a high-pass IIR filter of 0.16Hz to each participant's EEG data file. Then, for each EEG file, I applied a Fast Fourier Transform (FFT) with a Hamming window function to every

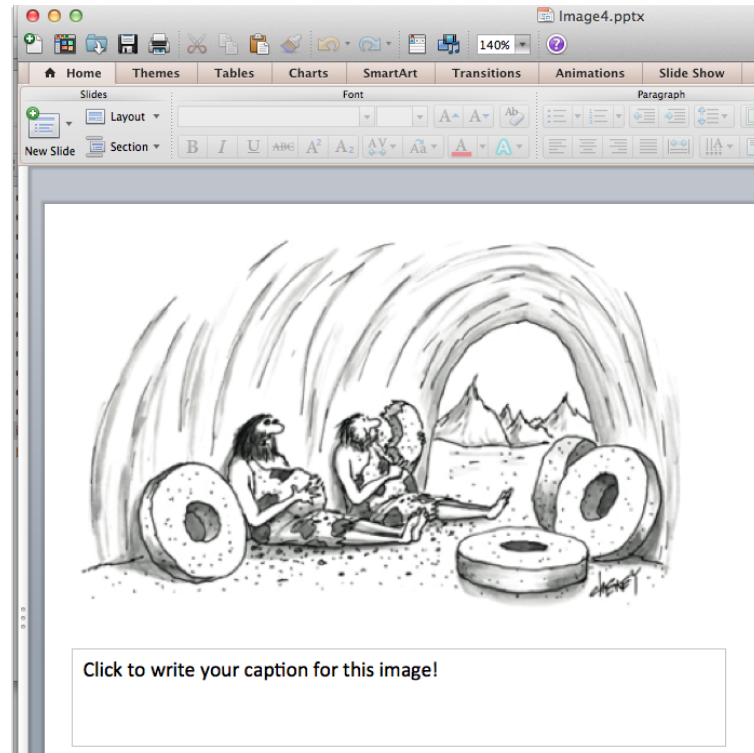


Figure 18: Example of an image that participants wrote a caption in response to.

two seconds of data. Since the EEG headset's sampling rate is 128Hz, I ran FFTs on every 256 samples. Within each FFT, I calculated average magnitude across 14 bands for every two seconds of data. The bands that I used in this study were: 2-4Hz, 4-6Hz, 6-8Hz, etc., through 28-30Hz.

It is important to note that in Study 1, I used the traditional neuroscience bands: alpha band (7.5-12.5Hz), low beta (12.5-25Hz), and high beta (25-30Hz). However, I selected a different approach this time for several reasons. First, these neuroscience bands are not concrete. The frequency for alpha, beta, etc. often varies across research papers. And finally, more bands would give more data to perform machine learning on.

5.3.2 Affectiva Q

The sensor data from the Affectiva Q were normalized between 0-1. This data includes: electrodermal activity (EDA) (i.e., a measure of eccrine sweat glands) and accelerometer values for X, Y, and Z. I extracted features from these normalized values by creating proportion bands: 0 to <0.2, 0.2 to <0.4, 0.4 to <0.6, 0.6 to <0.8, ≥ 0.8 . Specifically, I looked at the proportion of data that occurred within each of these five bands across a two-second time period, for all sensors.

5.3.3 Keyboard/Mouse Logger

I created a basic keyboard and mouse logger that would log the timestamp for every single keystroke or mouse movement. I transformed this data by summing the number of times that the mouse was dragged, moved, or clicked, and the number of key presses, within two seconds.

5.3.4 Self-Report Ratings

The self-reported creativity ratings from the ITMC 2.0 Reporting application were used as labeled data for each person's machine learning classifier. Specifically, I calculated the average creativity ratings for every two seconds of temporal self-reporting of creative experience, and these average ratings became labels in the classifier. In this study, I employed a machine learning classifier based on linear regression (rather than logistic regression or binary classification), therefore I did not have to discretize the creativity ratings. Rather, they remained as continuous variables on a scale of 0-100. The primary reason for not discretizing the creativity ratings is because par-

ticipants seemed to use the rating scale differently, with some participants not taking advantage of the full rating scale. This is likely a result of not giving participants a specific definition of creativity. Therefore, there was no cut off point for high and low creativity that would be appropriate across participants.

5.4 Machine Learning Procedure

I used WEKA [25] to create machine learning classifiers for every participant's sketching and writing tasks. For each person ($n=24$), I had two separate data sets, a Sketching Set and a Writing Set, for a total of 48 data sets. Within each data set, the labeled data was the person's temporal, self-reported ratings of creative experience, and the unlabeled data used for classification were the features extracted from the sensor packages previously described above in Sections 5.2.1 and 5.3. Feature extraction from these sensors resulted in 229 total features. In machine learning, too many features can result in a classifier being overfit, therefore feature selection can be used to reduce the data [1]. To do this, I used a Principle Components Analysis (PCA), which reduced the features in the classifier from 229 features to an average of 87 features ($SD=32.24$). In other words, the feature selection resulted in a different number of features for each participant.

The selected features (or PCA components) were then passed into a machine learning classifier to serve as the attributes data which would be classified based on the self-reported ratings of creative experience (i.e., the labeled data or ground truth). For machine learning, I employed a sequential minimal optimization algorithm, which utilizes a support vector machine for linear regression (e.g., 'SMOreg' in WEKA).

Each data set was divided into a Training Set and a Testing Set using 10-fold cross validation, and it was repeated for 5 iterations. Ten-fold cross-validation works by using 9 folds (or 90% of the data) as individual Training Sets and then using 1 fold (or 10% of the data) for a Testing Set. Each fold is comprised of randomly selected data. There were five iterations of 10-fold cross validation for each participant’s data set, and WEKA reported the average across across all five iterations.

Classification accuracy for SMOreg is based on correlation coefficients, since this classifier is based on linear regression, rather than logistic regression (i.e. binary classification). In interpreting these correlations, I used Cohen’s guidelines for correlation interpretation [14]. Cohen’s guidelines are: 0.20 (small effect size), 0.30 (medium effect size), and 0.50 (large effect size). This is an important distinction to make in contrast to machine learning in traditional computer science, since Cohen’s guidelines are directed towards the social sciences. For example, if we are using machine learning to create a classifier based on mouse clicks as the labeled data, we know concretely that a mouse click is a mouse click. However, in the case of self-reported data of creative experience, we must recognize that human error and subjectivity play a role in self-report, and this level of ambiguity in the labeled data must be considered. Thus, Cohen’s guidelines for the social sciences are appropriate.

5.5 Classification Results

5.5.1 Drawing and Writing

The correlation between self-reported creative experience and sensor data for the Sketching Set resulted in an average classification accuracy of $R=0.34$ ($SD=0.19$)

across all participants, which is a medium effect size. Specifically, there were five large effect sizes, (≥ 0.50), 11 medium effect sizes (≥ 0.30), and three small effect sizes (≥ 0.20). These correlations are summarized in Figure 19 and are listed in the Appendix in Table 15. Three participants had correlations below 0.20, and two participants had correlations below 0.10. The correlation average was computed by converting all correlations to Fisher's Z' , averaging the Z' values, and converting the Z' average back to a correlation. This is a standardized process for averaging correlations [59].

Similarly, correlations were found for self-reported creative experience and sensor data for the Writing Set, but the effect size was lower. In this case, the average classifier accuracy across all participants was a small effect size at $R=0.27$ ($SD=0.16$). There were two large effect sizes (≥ 0.50), ten medium effect sizes (≥ 0.30), and four small effect sizes (≥ 0.20). Six participants had correlations below 0.20, and two participants had correlations below 0.10. A summary is available in Figure 20 for a summary and are listed in the Appendix in Table 16.

5.5.2 Sensor Comparison

Since the above analysis resulted in promising classifier correlations, I wanted to explore which sensor package was the most effective for classifying creative experience. While a traditional machine learning approach would involve feature selection to find the variables that most accounted for the variance in the data, this does not tell us which sensor package performs the best. Perhaps a researcher does not have access to an accelerometer and wonders whether a keyboard/mouse logger will be sufficient.

It was my hope that by doing this level of analysis, it would allow researchers to better understand which sensors are necessary for doing this type of work. This level of analysis also allowed me to explore whether activity was a proxy for self-reporting creative experience by looking at the performance of classifiers based on activity sensors.

For each sensor package, I trained and tested an SMOreg classifier for that sensor alone, creating classifiers for four sensor packages: EEG (206 features), EDA (five features), Keyboard/Mouse Logger (three total features), and Accelerometer (15 total features). These accuracy results are in Table 12.

As shown in Table 12, there is no clear indication of one sensor being the total solution. However, it is evident that electrodermal activity (or EDA) did not contribute to the classification accuracy. As an additional check, I removed EDA and performed the same machine learning procedures for training and testing the Sketching Set and the Writing Set, and removing EDA did not change the accuracy for all sensors that is reported in Table 12.

For Sketching, both EEG ($R=0.24$, $SD=0.15$) and the Accelerometer ($R=0.22$, $SD=.018$) performed well, indicating that participants were not self-reporting exclusively based on on-screen activity. In the Writing Task, all sensors performed with fairly equivalent accuracy; however, EEG did perform the poorest at ($R=0.14$, $SD=0.11$), in contrast to the Logger ($R=0.19$, $SD=0.18$) and the Accelerometer ($R=0.18$, $SD=0.15$), which may offer some support that participants were likely self-reporting based on activity, in comparison to Sketching.

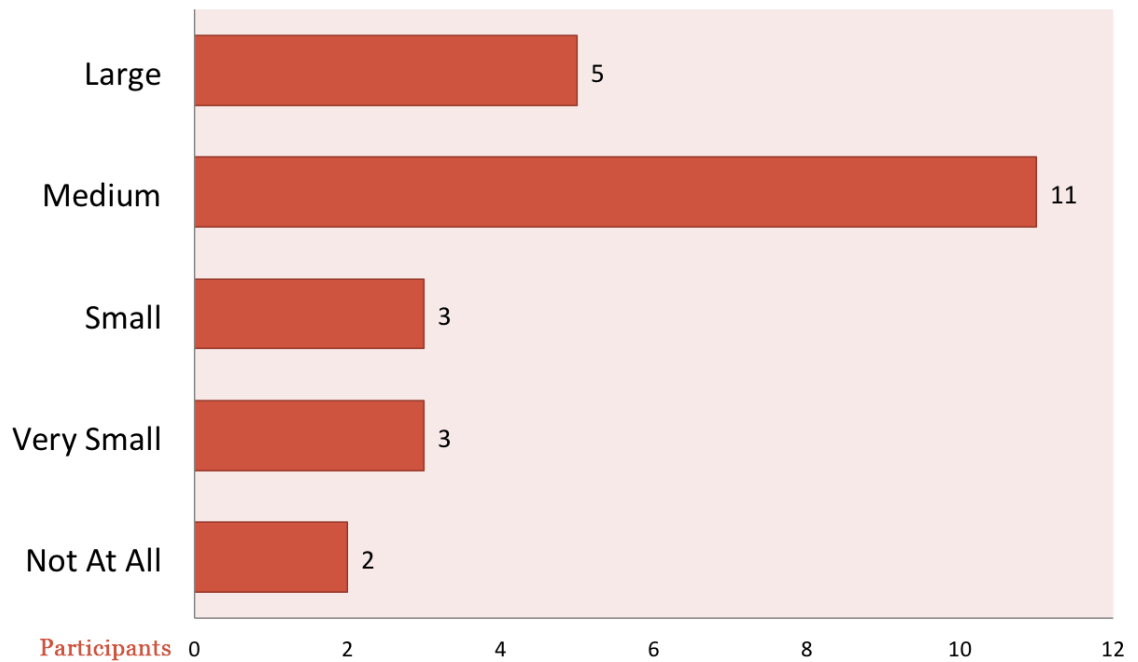


Figure 19: This chart summarizes the correlation effect sizes across participants for sketching. Sixteen of 24 participants performed with moderate to high accuracy.

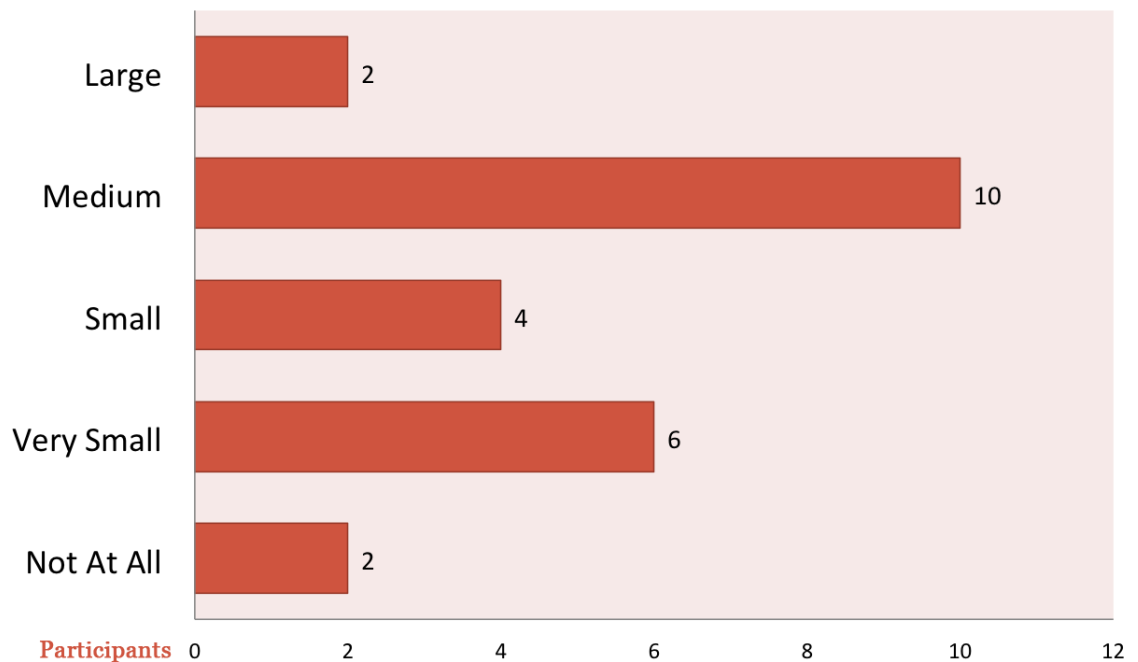


Figure 20: Similarly, this chart summarizes the correlation effect sizes across participants for writing. In this case, 12 out of 24 participants performed with moderate to high accuracy.

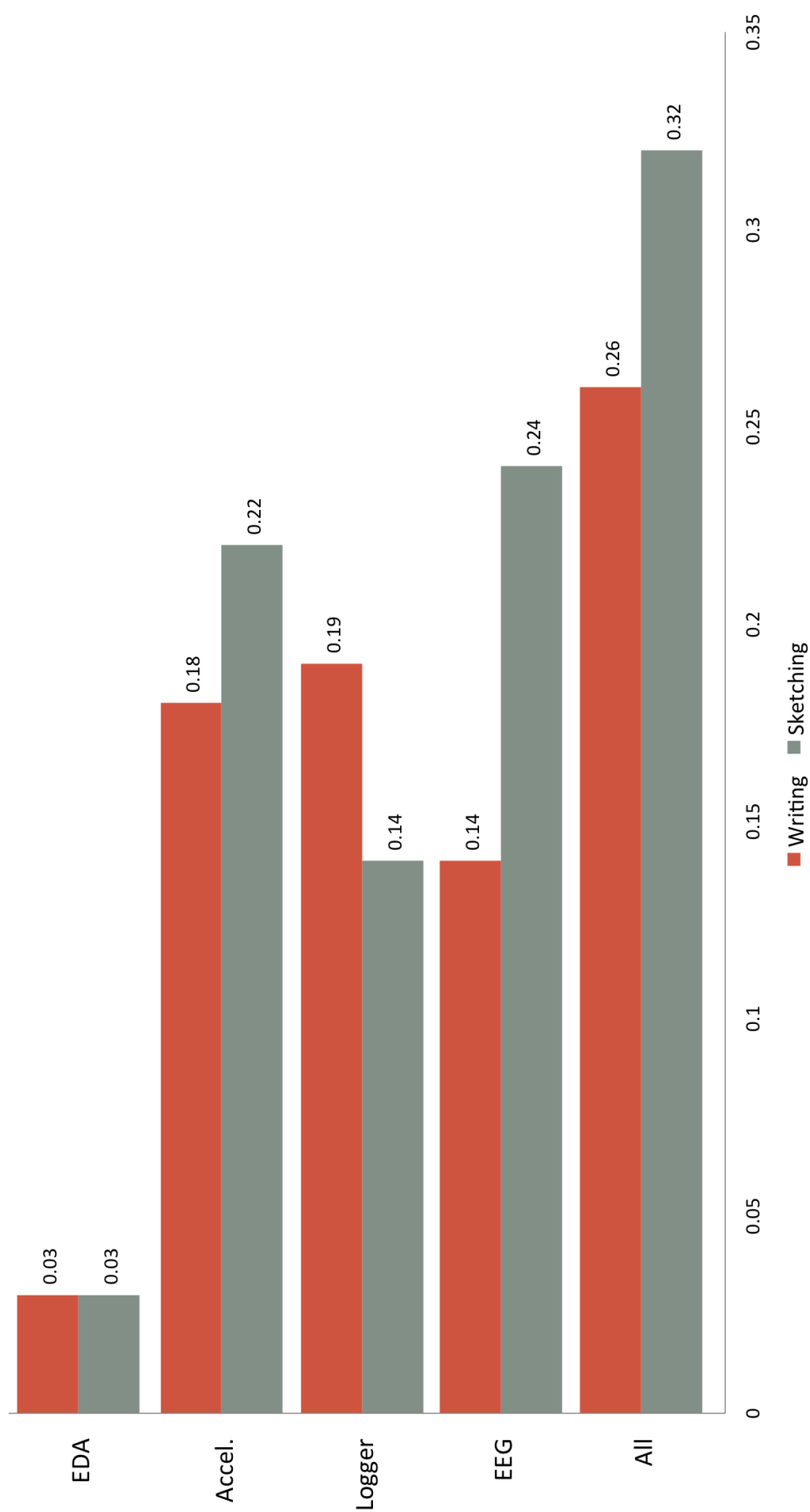


Figure 21: Average classification accuracy across five different machine learning classifiers. The EDA classifier performed poorly for both tasks. Activity classifiers (i.e., Accelerometer & Logger) performed best for Writing. EEG and Accelerometer performed best for Sketching.

Table 12: Classification accuracy of SMOreg for Sketching and Writing for all different sensor packages: EEG, Logger (i.e., Keyboard/Mouse logger), Accelerometer, and EDA. The accuracy of using all sensors is also displayed. **All* refers to using EEG, Logger, Accelerometer, and EDA in one classifier.

Avg. Correlation Coefficient (SD)					
<i>Task</i>	<i>All</i>	<i>EEG</i>	<i>Logger</i>	<i>Accel.</i>	<i>EDA</i>
Sketching	0.32 (0.19)	0.24 (0.15)	0.14 (0.13)	0.22 (0.18)	0.03 (0.07)
Writing	0.26 (0.15)	0.14 (0.11)	0.19 (0.18)	0.18 (0.15)	0.03 (0.05)

As a reminder, participants were allowed to use their own personal definition of creative experience when self-reporting their creativity. It is likely that the difference in sensor performance between Sketching and Writing results in participants using a different personal definition of creative experience to rate creativity in response to drawing, in contrast to rating linguistic creativity.

5.5.3 Test-Retest

As previously discussed in Section 5.2.2, there were 24 participants in this user study, and five participants were randomly selected to participate in the study a second time three weeks later. The purpose of having participants come back for a second session was to test whether a participant’s classifier from Session 1 could reliably be used in Session 2. Across these five participants, the test-retest for the Sketching task resulted in an average SMOreg accuracy of $R=0.35$ ($SD=0.04$) and for the Writing task, it resulted in an average SMOreg accuracy of $R=0.36$ ($SD=0.10$). These results suggest that a person’s classifier would be reliable over time, when applied to the same creative task.

Table 13: This table summarizes the correlation accuracies for a machine learning classifier that is based on combining Sketching+Writing data. The activity metrics (i.e. Accelerometer and Logger) were the most generalizable to both tasks.

Avg. Correlation Coefficient (SD)				
<i>All</i>	<i>EEG</i>	<i>EDA</i>	<i>Accelerometer</i>	<i>Logger</i>
0.39 (0.24)	0.28 (0.21)	0.02 (0.05)	0.34 (0.25)	0.31 (0.27)

5.5.4 Cross-Task Classification

I also investigated cross-task classification. In other words, was it possible to train on Sketching and then test on Writing? Given the results of the sensor comparison in Section 5.5.2, it seemed unlikely that a classifier for Sketching would be relevant to Writing and vice versa. To briefly examine cross-task classification, I created classifiers for several participants based on their Sketching data and then tested on their Writing data, and vice versa. As expected, this approach was not fruitful, and the classifiers were very weakly correlated.

However, I took a second approach to looking at the generalizability of classifiers based on tasks. For this approach, I took both data sets (Sketching and Drawing) for every single participant, and I combined them into one file. Then, I performed 10-fold cross validation across 5 iterations using the SMOreg classifier. This means that both the training set and the test set was randomly composed of data from Sketching and from Drawing. This approach was very successful, and it resulted in an average classification of $R=0.39$ ($SD=0.24$).

An interesting outcome of this work is that the overall classification accuracy was actually higher than the overall classification accuracy for Sketching alone: $R=0.39$

(Sketching & Writing combined) compared to $R=0.32$ (Sketching alone). This is mostly due to the fact that by combining Sketching and Writing into one data set, more training data was available to the classifier.

In order to understand what sensor packages performed the best for this classifier based on Sketching+Writing, I also created classifiers based on sensor packages. In comparing classifiers across a Sketching+Writing data set, the activity sensors performed the best. These results are: EEG ($R=0.28$, $SD=0.21$), EDA ($R=0.02$, $SD=0.05$), Accelerometer ($R=0.34$, $SD=0.25$), and Keyboard/Mouse Logger ($R=0.31$, $SD=0.27$). These results are summarized in Table 13.

These results indicate that it is possible to create a generalized classifier based on Sketching+Writing; however, it appears that this generalized classifier is primarily based on activity metrics and may be more about creative execution.

5.5.5 Cross-User Classification

I also investigated cross-user classification: the extent to which one participant's classifier could be used on a different participant. To investigate this, I selected four participants with moderate to high classification accuracy for Sketching. I combined three of the participants' data for a training set, and then I tested on the fourth participant. This process of training on three participants and testing on a fourth participants was also repeated for the Writing task.

Unfortunately, it was not possible to create a classifier based on cross-user data. In other words, there was no correlation for the SMOreg classifier. This was an expected result, given that: (1) participants were allowed to use their own personal definition

Table 14: Results of the post experiment survey.

Question	Avg (SD)	Rating Scale
Please rate how creative you felt overall while sketching during this study.	3.17 (0.65)	(1)Not At All Creative, (5)Extremely Creative
Please rate how creative you felt overall while writing captions during this study.	3.48 (1.12)	
Please rate how creative you typically feel when sketching.	3.65 (0.93)	
Please rate how creative you typically feel when writing creatively.	3.57 (0.99)	
Please rate how comfortable you felt when asked to self-report your creativity.	3.48 (1.31)	(1)Not At All Comfortable, (5)Extremely Comfortable
Please rate how accurate you felt when asked to self-report your creativity.	3.58 (1.12)	(1)Not At All Accurate, (5)Extremely Accurate
Please rate how easy or difficult it was to use the video interface for reporting your creativity.	3.91 (1.20)	(1)Not At All Easy To Use, (5)Extremely Easy To Use

of what it meant to be creative, and (2) in related work, Grimes et al. were not able to develop an EEG classifier for cross-user classification based on mental workload, which is less ambiguous than a study of creative experience [23].

5.5.6 Other Surveys

As previously discussed, participants also took the Creative Behavior Inventory (CBI) and also filled out a Post Experiment Survey. To score the CBI, I summed all 77-questions and then normalized them between 0-100, and the average score on the CBI was 45.33 (SD=28.24). The results of the Post Experiment Survey are available in Table 14. Neither the results from the CBI nor the Post Experiment Survey were useful in the analysis of this study.

5.6 Video Analysis

While this research is focused on a quantitative approach to evaluating creativity support tools, I also viewed several participants' screen recording videos, while playing back their ITMC ratings, inside the ITMC Reporting application. In viewing this data, I was interested in seeing whether there were patterns in how participants self-reported their creative experience, as well as looking at the temporal shape of the data. Since this was a crude video analysis, I only selected a subset of participants' data to view. I selected two participants with the most accurate classifiers (i.e., one best from Sketching and one best from Writing), and similarly, two participants with the least accurate classifiers.

For the participant with the lowest classification accuracy for Sketching, it seemed that, for the most part, the participant only moved his/her creative experience up or down at the start of each task. In contrast, the participant with the lowest classification accuracy for Writing, there were no clear patterns. Sometimes the participant left the creative experience rating at the same level throughout a task; other times it was continuously decreasing or increasing through a task.

There were more similarities for the participants with the best classification accuracy for Sketching and Writing. For both of these participants, the self-reported ratings of creative experience were overall relatively low, taking up the bottom half of the scale; however, both participants had several tasks with very clear creative peaks.

In future ITMC studies, it may be beneficial to use video analysis as a post-hoc tool, in conjunction with interviews. In other words, it would be interesting to still

have participants retrospectively self-report their creativity. Afterwards, interviews would be conducted to show participants their data and ask them why their creative experience went up or down. This would not only be beneficial in understanding *why* someone’s creative experience went up or down, but it may also help identify errors in their self-report.

5.7 Discussion

In both ITMC Study 1 and 2, I demonstrated that creative experience can be predicted with moderate accuracy, and furthermore, my results in Study 2 indicate the potential for a classifier to work across sessions and across tasks. Given that physiometrics and brain-computer interaction is largely in its infancy, these moderate results in predicting creative experience using sensor data are important and promising.

In moving forward with this line of research, it is critical to reduce the measurement error in self-report in order to do machine learning classification, and there are two primary ways in which this can be done: (1) Providing a definition of creative experience to participants, and (2) Conducting deeper qualitative user studies in order to find why certain features in a creativity support tool promote or hinder creativity. In the first case, researchers should proceed with caution, as providing definitions of creative experience could be difficult for users and typically decreases the reliability. Therefore, in providing any definition of creativity, the user should be given a training period to become accustomed to working with the definition. In the second case, qualitative studies should involve in-depth interviews with participants using post-hoc video protocols.

5.8 Future Work

The results of both ITMC Study 1 and Study 2 clearly demonstrate that there is some true signal in both the electroencephalography (EEG) data, as well as the activity metrics, that is useful for capturing measures of creative experience. However, given that this research in employing sensor data to capture creative experience is largely still in its infancy, there is much more research to be conducted in this area.

In future work, it is important to design studies that further isolate the phenomenon of creative experience from confounding variables. For example, it would be beneficial to design user studies that allow us to measure the extent to which the signal is capturing creative experience versus motor movements. Such studies that would be beneficial in future work include:

- Studies in which people perform both dull, mundane tasks, in addition to tasks that are explicitly creative and engaging. Ideally, a signal should be able to distinguish between dull tasks versus creative tasks.
- Studies in which people perform derivative tasks, like reproducing an existing sketch, in addition to creating original art. Similar to the study above, there should be differences in the signal for reproducing existing work versus creating original work.

CHAPTER 6: CONCLUSION

6.1 Dissertation Contributions

As discussed in Chapter 1, there are no obvious quantitative measurements for evaluating creativity support tools [57], and furthermore, it is even more challenging to define and measure creativity itself (Chapter 2). To address these challenges, my research has contributed two computational methodologies for evaluating creativity support tools: a psychometric tool called the Creativity Support Index (CSI) and a physiological sensor data approach to measuring ‘in-the-moment-creativity’ (ITMC).

6.1.1 Creativity Support Index

The Creativity Support Index (Chapter 3) is a psychometrically developed and validated survey that is intended to be administered to participants after using a creativity support tool. The CSI is comprised of six factors: Collaboration, Enjoyment, Exploration, Expressiveness, Immersion, and Results Worth Effort. There are two agreement statements per factor for a total of 12 agreement statements. There is also a paired factor comparison section, in which each factor is compared against every other factor for a total of 15 comparisons. These paired comparisons are used to weigh the CSI score according to the factors that are the most relevant to the creative task. Therefore, the CSI produces a weighted score out of 100, where a higher score indicates better creativity support.

6.1.2 ‘In-The-Moment-Creativity’

The goal of this research was to detect *when* a person is experiencing high creativity while using a creativity support tool. In other words, how can we evaluate a tool’s ability to support creativity, if we are not able to detect when someone is actually being creative? To this end, this dissertation investigated ‘in-the-moment-creativity’ (ITMC) using a computational, machine learning approach for detecting moments of high creative experience with physiological sensor data and self-reported ratings of creative experience.

Within this ITMC research, there are several dissertation contributions. First, this dissertation conceptualized ‘in-the-moment-creativity,’ and it presented the first user studies that investigated temporal self-reporting of creative experience. Second, this dissertation presented studies that found moderate accuracy for detecting high creative experience using physiological sensor data and ratings of creative experience (Chapters 4, 5). These results indicate that not only was it possible for participants to temporally self-report creative experience, but also that there was a relationship between physiological measurements and self-reported ratings of creative experience. And finally, this dissertation also presented two iterations of a custom video application for self-reporting ITMC.

The other questions explored in this dissertation were: (1) whether activity measurements could be a proxy for studying creative experience, and (2) how generalizable a person’s machine learning classifier was. The first point was focused on investigating whether participants were self-reporting high creative experience only when there

was lots of on-screen activity in their screen recording videos. The results of ITMC Study 2 in Chapter 5 found that in the Sketching task, both EEG and activity metrics were crucial in the machine learning classifier. However, for the Writing task, the machine learning classifier was primarily comprised of activity metrics. These results indicate that in the Sketching task, participants were self-reporting high creativity not only when there was lots of on-screen activity, but also when they were feeling creative. However, for Writing, it indicates that they were self-reporting high creativity primarily when there was lots of on-screen activity. More than likely, these results are based on participants having different conceptual definitions of creativity between Sketching and Writing.

The second point was focused on the generalizability of the classifier. Would a person's Sketching classifier work on a Writing task? Would one person's classifier work on someone else's data? Would a person's classifier work across time? These questions were addressed in Chapter 5. To summarize, this dissertation found that cross-task classifier was not possible; however, a generalized classifier could be developed by combining a person's Sketching and Writing data. This work also found moderate accuracy for test-retest, which means that a user's classifier works across multiple sessions (e.g., or across time). However, cross-user classification was not possible. This result was expected, given that people were allowed to use their own definition of creative experience, and it was these self-reported ratings of creative experience which were used as labeled data (or ground truth) in the classifier.

6.2 Future Research

As discussed throughout this dissertation, there are many future research directions for studying ITMC, and this section summarizes them.

6.2.1 Adaptive Interfaces

The area of future research that interests me the most is using ITMC detection as input into adaptive interfaces. In other words, maybe the system recognizes that the user is in a state of creative flow, and it turns off all notifications to avoid disrupting the user. Or, perhaps it recognizes that a person is not being creative, and it cues classical music to help elevate the user's creativity. This brings up another interesting point in that ITMC detection could also allow us to investigate environmental or contextual impacts, such as the effect of background music on the creative process.

However, in order to make adaptive interfaces for creative experience possible, the first line of research is to reduce measurement error in self-reported ratings of creative experience. Since the labeled data in the machine learning classifiers are based on self-reported ratings of creative experiences, it is critical to investigate ways to reduce the ambiguity in the labeled data. As I suggested in Chapter 5, a first step in reducing measurement error is to provide a definition of creative experience to the user, or at least scope it down. While Amabile [5] reported that allowing participants to choose their own definition of creativity increases the reliability of creativity ratings, one possibility is to give users an actual definition of creative experience and then provide them with a training session to get accustomed to the formal definition.

6.2.2 Self-Reported Ratings of Creative Experience

Clearly, improving the reliability of self-reported ratings of creative experience is essential for future research, and this can also be studied without physiological sensors. In other words, ITMC reliability can be studied in experiments in which participants still retrospectively self-report their creative experience; however, this would happen in conjunction with deep, qualitative user studies. Participants would not only be asked *when* they were experiencing creativity, but also *why* certain features in a CST promoted or hindered their creativity. Not only would this process provide us with a better understanding of the CST being studied, but it should also help participants discover errors in their self-report.

Finally, it is worth noting that formal psychometrics methodologies would not work for measuring the reliability of this time series data. Typically, this would be studied in a test-retest situation. However, the task varies from task 1 to task 2; so it would not be possible to measure error with this procedure.

6.2.3 Ideation vs. Execution

There is also a need to devise experiments that attempt to separate ideation and execution, as it is possible that they have different sensor data signatures. In other words, sensor data in a High Creative Experience may have different signature for creative execution than for creative ideation.

Because the first ITMC study involved participants coming to the study with pre-conceived ideas of what to sketch, it is possible that the creative experience being measured was mostly executional creativity. I wanted to be sure to capture both

ideation and execution in Study 2, which is why the study was devised as a series of short tasks with prompts. I believe that both ideation and execution were captured in Study 2; however, it was not possible to tease these out and build separate classifiers for them. In reviewing the participants' screen recording videos, it seemed that ideation and execution were highly intertwined. For example, when writing captions in response to cartoon prompts, participants often slowly typed their responses, and often changed their mind about what they were writing. Even if a few moments were able to be identified as being clearly ideation or clearly execution, it would not be enough data for machine learning classification.

There are two possibilities for exploring the differences between ideation and execution in the future. The most methodological way would be to design experiments with creative tasks that specifically separate ideation and execution. For example, in the case of creative writing, this has been done in the past: segmenting the experiment to where participants are first instructed to spend some time thinking about a creative story that they want to write, and then in the other half of the experiment, allowing them to execute that story [48]. Another possible way to at least control for differences between creative ideation and creative execution is by providing participants a definition of creative experience. In other words, participants could be provided a definition that ignores either ideation or execution.

REFERENCES

- [1] E. Alpaydin. *Introduction to Machine Learning*. MIT Press, Cambridge, MA, 2 edition, 2009.
- [2] T. M. Amabile. Children’s Artistic Creativity: Detrimental effects of competition in a field setting. *Personality and Social Psychology Bulletin*, 8(3):573–578, 1982.
- [3] T. M. Amabile. *Growing Up Creative*. Creative Education Foundation, 1992.
- [4] T. M. Amabile. Beyond Talent: John Irving and the Passionate Craft of Creativity. *American Psychologist*, 56(4):333–336, Apr. 2001.
- [5] T. M. Amabile, P. Goldfarb, and S. C. Brackfield. Social influences on creativity: Evaluation, coaction, and surveillance. *Creativity Research Journal*, 3:6–21, 1990.
- [6] R. Beghetto. Toward a broader conception of creativity: A case for “mini-c” creativity. *Psychology of Aesthetics*, 2007.
- [7] M. A. Boden. *The creative mind: Myths and Mechanisms*. Psychology Press, 2004.
- [8] L. Candy. Evaluating Creativity. In *Creativity and Rationale: Enhancing Human Experience by Design*, pages 57–84. Springer-Verlag.
- [9] L. Candy. Constraints and creativity in the digital arts. *Leonardo*, 40:366–367, 2007.
- [10] E. A. Carroll and C. Latulipe. The Creativity Support Index. In *ACM CHI 2008 Extended Abstracts*, Apr. 2009.
- [11] E. A. Carroll and C. Latulipe. Capturing ‘In the Moment’ Creativity Through Data Triangulation. In *ACM Creativity & Cognition ’11 Extended Abstracts*, Nov. 2011.
- [12] E. A. Carroll and C. Latulipe. Triangulating the Personal Creative Experience: Self-Report, External Judgments, and Physiology. In *Proceedings of Graphics Interface 2012*, pages 53–60, 2012.
- [13] E. A. Carroll, C. Latulipe, R. Fung, and M. Terry. Creativity Factor Evaluation: Towards a Standardized Survey Metric for Creativity Support. In *Proceedings of ACM Creativity & Cognition 2009*. ACM Press, 2009.
- [14] J. Cohen. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2 edition, 1988.
- [15] M. Csikszentmihalyi. *Creativity: Flow and the Psychology of Discovery and Invention*. Harper Perennial, 1997.

- [16] G. Davis and T. L. Belcher. How shall creativity be measured? Torrance tests, RAT, Alpha Biographical, and IQ. *Journal of Creative Behavior*, 5:153–161, 1971.
- [17] A. Delorme, T. Sejnowski, and S. Makeig. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *Neuroimage*, pages 1143–1149, 2007.
- [18] S. P. Dow, A. Glassco, J. Kass, M. Schwarz, D. L. Schwartz, and S. R. Klemmer. Parallel Prototyping Leads to Better Design Results, More Divergence, and Increased Self-Efficacy. *ACM Transactions on Computer-Human Interaction*, 17(4):18–24, Dec. 2010.
- [19] P. Ekman and E. L. Rosenburg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FAC)*. Oxford University Press, New York, 2004.
- [20] R. Fung, E. Lank, M. Terry, and C. Latulipe. Kinematic templates: End-user tools for content-relative cursor manipulations. In *Proceedings of ACM UIST '08*, 2008.
- [21] J. W. Getzels and P. W. Jackson. *Creativity and intelligence: Explorations with gifted students*. Wiley, New York, 1962.
- [22] B. Gonzalez and C. Latulipe. BiCEP: Bimanual Color Exploration Plugin. In *ACM CHI '11 Extended Abstracts*, pages 1483–1488. ACM Press, May 2011.
- [23] D. Grimes, D. Tan, S. Hudson, P. Shenoy, and R. Rao. Feasibility and pragmatics of classifying working memory load with an electroencephalograph. *Proceedings of ACM CHI 2008*, 2008.
- [24] Z. Guan, S. Lee, E. Cuddihy, and J. Ramey. The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proceedings of ACM CHI '06*, pages 1253–1262, New York, New York, USA, 2006. ACM Press.
- [25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), 2009.
- [26] S. G. Hart. NASA Task Load Index (NASA TLX); 20 Years Later. In *Human Factors and Ergonomics Society Annual Meeting Proceedings*, pages 904–908. Human Factors and Ergonomics Society, 2006.
- [27] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload*, 1988.
- [28] T. Hewett, M. Czerwinski, M. Terry, J. Nunamaker, L. Candy, B. Kulers, and E. Sylvan. Creativity Support Tool Evaluation Methods and Metrics. In *NSF Workshop Report on Creativity Support Tools*, pages 10–24, 2005.

- [29] D. Hocevar. Intelligence, divergent thinking, and creativity. *Intelligence*, 4:25–40, 1980.
- [30] D. Hocevar. Measurement of Creativity: Review and Critique. *Journal of Personality Assessment*, 45(5):450, 1981.
- [31] M. J. A. Howe. *Genius Explained*. Cambridge University Press, United Kingdom, 2000.
- [32] W. James. *The Principles of Psychology (2 vols.)*. Henry Holt (Reprinted Bristol: Thoemmes Press: 1999), 1890.
- [33] H. Kaiser. The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20:141–151, 1960.
- [34] M. Karlins, C. Schuerhoff, and M. Kaplan. Some factors related to architectural creativity in graduating architecture students. *Journal of General Psychology*, 81:203–215, 1969.
- [35] J. C. Kaufman and R. A. Beghetto. Beyond big and little: The four c model of creativity. *Review of General Psychology*, 13(1):1–12, 2009.
- [36] A. Kerne, C. Latulipe, A. Webb, and E. A. Carroll. Workshop on Evaluation Methods for Creativity Support Environments. In *Fifth International Conference on Design Computing and Cognition (DCC) 2012*, College Station, Texas, 2011.
- [37] A. Kerne and S. M. Smith. The Information Discovery Framework. In *Proceedings of DIS '04*, pages 357–360. ACM Press, 2004.
- [38] A. Kerne, S. M. Smith, E. Koh, H. Choi, and R. Graeber. An Experimental Method for Measuring the Emergence of New Ideas in Information Discovery. *International Journal of Human-Computer Interaction*, 24(5):460–477, June 2008.
- [39] M. J. Kim and M. L. Maher. Comparison of designers using a tangible user interface and a graphical user interface and the impact on spatial cognition. Technical report, 2005.
- [40] P. Lang. The emotion probe: Studies of motivation and attention. *American Psychologist*, 1995.
- [41] C. Latulipe, I. Bell, C. Clarke, and C. Kaplan. symTone: Two-Handed Manipulation of Tone Reproduction Curves. In *Proceedings of GI '06*, pages 9–16, 2006.
- [42] C. Latulipe, E. A. Carroll, and D. Lottridge. Love, Hate, Arousal, and Engagement: Exploring Audience Responses to Performing Arts. In *Proceedings of ACM CHI 2011*, pages 1845–1854. ACM Press, May 2011.

- [43] C. Latulipe, S. Mann, C. L. A. Clarke, and C. S. Kaplan. SymSpline: Symmetric two-handed spline manipulation. In *Proceedings of ACM CHI 2006*, pages 349–358, 2006.
- [44] M. Maceli and M. Atwood. From human factors to human actor to human crafters. In *Proceedings of the 2011 iConference*, pages 98–105, 2011.
- [45] M. L. Maher. Computational and Collective Creativity: Who’s being creative? In *Proceedings of the Third International Conference on Computational Creativity*, pages 67–71, 2012.
- [46] M. L. Maher and D. H. Fisher. Using AI to Evaluate Creative Designs. In *Proceedings of the International Conference on Creative Design*, 2012.
- [47] R. L. Mandryk and K. M. Inkpen. Physiological Indicators for the Evaluation of Co-Located Collaborative Play. In *Proceedings of CSCW 2004*, pages 102–111, 2004.
- [48] C. Martindale. *Biological Bases of Creativity*. Cambridge University Press, 1999.
- [49] C. Martindale and N. Hasenfus. EEG differences as a function of creativity, stage of the creative process, and effort to be original. *Biological Psychology*, 6:157–167, 1978.
- [50] M. Mikhail, K. El-Ayat, R. El Kaliouby, J. Coan, and J. J. B. Allen. Emotion detection using noisy EEG data. In *Proceedings of the 1st Augmented Human International Conference*, 2010.
- [51] D. Norton, D. Heath, and D. Ventura. An Artistic Dialogue with the Artificial. In *Proceedings of ACM Creativity and Cognition 2011*, pages 31–40, 2011.
- [52] R. W. Picard. *Affective computing*. MIT Press, Cambridge, MA, USA, 1997.
- [53] J. A. Plucker and J. S. Renzulli. *Psychometric approaches to the study of human creativity*. Cambridge University Press, 1999.
- [54] M. Z. Poh, N. C. Swenson, and R. W. Picard. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE Transactions on Biomedical Engineering*, 57(5):1243–1252, May 2010.
- [55] J. Read, S. MacFarlane, and C. Casey. Endurability, engagement, and expectations: Measuring children’s fun. In *IDC ’02*. Shake Publishing, 2002.
- [56] K. Rubin, G. Fein, and B. Vandenburg. Play. In P. Mussen and E. Hetherington, editors, *Handbook of Child Psychology*, pages 693–774. John Wiley and Sons, Inc., Wiley, New York, 1983.
- [57] B. Shneiderman. Creativity support tools: accelerating discovery and innovation. *Communications of the ACM*, 50(12):20–32, Dec. 2007.

- [58] B. Shneiderman, G. Fischer, and M. Czerwinski. Creativity support tools: Report from a US National Science Foundation sponsored workshop. *International Journal of Human-Computer Interaction*, 20:61–77, 2006.
- [59] N. C. Silver and W. P. Dunlap. Averaging correlation coefficients: Should Fisher’s z transformation be used? *Journal of Applied Psychology*, 72(1), 1987.
- [60] V. Singh, C. Latulipe, E. A. Carroll, and D. Lottridge. The Choreographer’s Notebook – A video annotation system for dancers and choreographers. In *Proceedings of ACM Creativity & Cognition ’11*, pages 197–206, Atlanta, GA, Nov. 2011.
- [61] E. P. Torrance. *Torrance Tests of Creative Thinking: Norms-technical manual*. Personnel Press/Ginn, New York, NY, 1974.

APPENDIX A: CREATIVITY SUPPORT INDEX

Creativity Support Index:
Agreement Statements – Page 1

Please rate your agreement with the following statements by placing a check mark in the corresponding box.

1. I was satisfied with what I got out of the system or tool.

(Highly
Disagree)

--	--	--	--	--	--	--	--	--	--

(Highly
Agree)

2. It was easy for me to explore many different ideas, options, designs, or outcomes, using this system or tool.

(Highly
Disagree)

--	--	--	--	--	--	--	--	--	--

(Highly
Agree)

☐ N/A

3. The system or tool allowed other people to work with me easily.

(Highly
Disagree)

--	--	--	--	--	--	--	--	--	--

(Highly
Agree)

4. I would be happy to use this system or tool on a regular basis.

(Highly
Disagree)

--	--	--	--	--	--	--	--	--	--

(Highly
Agree)

5. I was able to be very creative while doing the activity inside this system or tool.

(Highly
Disagree)

--	--	--	--	--	--	--	--	--	--

(Highly
Agree)

6. My attention was fully tuned to the activity, and I forgot about the system or tool that I was using.

(Highly
Disagree)

--	--	--	--	--	--	--	--	--	--

(Highly
Agree)

Creativity Support Index:
Agreement Statements – Page 2

Please rate your agreement with the following statements by placing a check mark in the corresponding box.

7. I enjoyed using this system or tool.

(Highly
Disagree)

--	--	--	--	--	--	--	--	--	--

(Highly
Agree)

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

(Highly
Disagree)

--	--	--	--	--	--	--	--	--	--

(Highly
Agree)

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

(Highly
Disagree)

--	--	--	--	--	--	--	--	--	--

(Highly
Agree)

10. The system or tool allowed me to be very expressive.

(Highly
Disagree)

--	--	--	--	--	--	--	--	--	--

(Highly
Agree)

11. It was really easy to share ideas and designs with other people inside this system or tool.

☐ N/A

(Highly
Disagree)

--	--	--	--	--	--	--	--	--	--

(Highly
Agree)

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

(Highly
Disagree)

--	--	--	--	--	--	--	--	--	--

(Highly
Agree)

Creativity Support Index:
Paired Comparison Section – Page 3

For each pair, please indicate your response to the following statement:

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

1.	<input type="checkbox"/> Explore many different ideas, outcomes, or possibilities	<input type="checkbox"/> Work with other people
2.	<input type="checkbox"/> Be creative and expressive	<input type="checkbox"/> Produce results that are worth the effort I put in
3.	<input type="checkbox"/> Enjoy using the system or tool	<input type="checkbox"/> Become immersed in the activity
4.	<input type="checkbox"/> Become immersed in the activity	<input type="checkbox"/> Produce results that are worth the effort I put in
5.	<input type="checkbox"/> Work with other people	<input type="checkbox"/> Enjoy using the system or tool
6.	<input type="checkbox"/> Produce results that are worth the effort I put in	<input type="checkbox"/> Explore many different ideas, outcomes, or possibilities
7.	<input type="checkbox"/> Be creative and expressive	<input type="checkbox"/> Become immersed in the activity
8.	<input type="checkbox"/> Work with other people	<input type="checkbox"/> Produce results that are worth the effort I put in
9.	<input type="checkbox"/> Be creative and expressive	<input type="checkbox"/> Enjoy using the system or tool
10.	<input type="checkbox"/> Explore many different ideas, outcomes, or possibilities	<input type="checkbox"/> Become immersed in the activity
11.	<input type="checkbox"/> Work with other people	<input type="checkbox"/> Be creative and expressive
12.	<input type="checkbox"/> Produce results that are worth the effort I put in	<input type="checkbox"/> Enjoy using the system or tool
13.	<input type="checkbox"/> Explore many different ideas, outcomes, or possibilities	<input type="checkbox"/> Be creative and expressive
14.	<input type="checkbox"/> Work with other people	<input type="checkbox"/> Become immersed in the activity
15.	<input type="checkbox"/> Explore many different ideas, outcomes, or possibilities	<input type="checkbox"/> Enjoy using the system or tool

Scoring Key

Results Worth Effort

1. I was satisfied with what I got out of the system or tool.

(Highly
Disagree)

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

(Highly
Agree)

Exploration

2. It was easy for me to explore many different ideas, options, designs, or outcomes, using this system or tool.

(Highly
Disagree)

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

(Highly
Agree)

Collaboration

☐ N/A (1)

3. The system or tool allowed other people to work with me easily.

(Highly
Disagree)

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

(Highly
Agree)

Enjoyment

4. I would be happy to use this system or tool on a regular basis.

(Highly
Disagree)

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

(Highly
Agree)

Expressiveness

5. I was able to be very creative while doing the activity inside this system or tool.

(Highly
Disagree)

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

(Highly
Agree)

Immersion

6. My attention was fully tuned to the activity, and I forgot about the system or tool that I was using.

(Highly
Disagree)

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

(Highly
Agree)

Scoring Key

Enjoyment

7. I enjoyed using this system or tool.

(Highly
Disagree)

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

(Highly
Agree)

Exploration

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

(Highly
Disagree)

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

(Highly
Agree)

Results Worth Effort

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

(Highly
Disagree)

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

(Highly
Agree)

Expressiveness

10. The system or tool allowed me to be very expressive.

(Highly
Disagree)

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

(Highly
Agree)

Collaboration

11. It was really easy to share ideas and designs with other people inside this system or tool.

☐ N/A (1)

(Highly
Disagree)

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

(Highly
Agree)

Immersion

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

(Highly
Disagree)

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

(Highly
Agree)

Scoring Key

1.	<input type="checkbox"/> Explore many different ideas, outcomes, or possibilities Exploration	<input type="checkbox"/> Work with other people Collaboration
2.	<input type="checkbox"/> Be creative and expressive Expressiveness	<input type="checkbox"/> Produce results that are worth the effort I put in Results Worth Effort
3.	<input type="checkbox"/> Enjoy using the system or tool Enjoyment	<input type="checkbox"/> Become immersed in the activity Immersion
4.	<input type="checkbox"/> Become immersed in the activity Immersion	<input type="checkbox"/> Produce results that are worth the effort I put in Results Worth Effort
5.	<input type="checkbox"/> Work with other people Collaboration	<input type="checkbox"/> Enjoy using the system or tool Enjoyment
6.	<input type="checkbox"/> Produce results that are worth the effort I put in Results Worth Effort	<input type="checkbox"/> Explore many different ideas, outcomes, or possibilities Exploration
7.	<input type="checkbox"/> Be creative and expressive Expressiveness	<input type="checkbox"/> Become immersed in the activity Immersion
8.	<input type="checkbox"/> Work with other people Collaboration	<input type="checkbox"/> Produce results that are worth the effort I put in Results Worth Effort
9.	<input type="checkbox"/> Be creative and expressive Expressiveness	<input type="checkbox"/> Enjoy using the system or tool Enjoyment
10.	<input type="checkbox"/> Explore many different ideas, outcomes, or possibilities Exploration	<input type="checkbox"/> Become immersed in the activity Immersion
11.	<input type="checkbox"/> Work with other people Collaboration	<input type="checkbox"/> Be creative and expressive Expressiveness
12.	<input type="checkbox"/> Produce results that are worth the effort I put in Results Worth Effort	<input type="checkbox"/> Enjoy using the system or tool Enjoyment
13.	<input type="checkbox"/> Explore many different ideas, outcomes, or possibilities Exploration	<input type="checkbox"/> Be creative and expressive Expressiveness
14.	<input type="checkbox"/> Work with other people Collaboration	<input type="checkbox"/> Become immersed in the activity Immersion
15.	<input type="checkbox"/> Explore many different ideas, outcomes, or possibilities Exploration	<input type="checkbox"/> Enjoy using the system or tool Enjoyment

Scoring the CSI Manually

Step 1) Compute the Scale Totals: Enter the participant's score for each of the six scales and sum them.

Results Worth Effort: Q1 (_____) + Q9 (_____) = _____

Exploration: Q2 (_____) + Q8 (_____) = _____

Collaboration: Q3 (_____) + Q11 (_____) = _____

Enjoyment: Q4 (_____) + Q7 (_____) = _____

Expressiveness: Q5 (_____) + Q10 (_____) = _____

Immersion: Q6 (_____) + Q12 (_____) = _____

Step 2) Count the Paired Comparisons: Count the total times that each factor was selected in the paired comparison section (pg.3 of the CSI).

Results Worth Effort: _____

Exploration: _____

Collaboration: _____

Enjoyment: _____

Expressiveness: _____

Immersion: _____

Step 3) Compute each Scale Score: For each scale, multiple the Scale Total (Step 1) times the count (Step 2).

(ResultsWorthEffort_{Total} * ResultsWorthEffort_{Count}): _____

(Exploration_{Total} * Exploration_{Count}): _____

(Collaboration_{Total} * Collaboration_{Count}): _____

(Enjoyment_{Total} * Enjoyment_{Count}): _____

(Expressiveness_{Total} * Expressiveness_{Count}): _____

(Immersion_{Total} * Immersion_{Count}): _____

Step 4) Compute the overall CSI Score: Sum each Scale Score (Step 3) and then divide by 3.0

$$\text{CSI Score} = \left[\text{ResultsWorthEffort}_{\text{Scale}} + \text{Exploration}_{\text{Scale}} + \text{Collaboration}_{\text{Scale}} + \text{Enjoyment}_{\text{Scale}} + \text{Expressiveness}_{\text{Scale}} + \text{Immersion}_{\text{Scale}} \right] / 3.0$$

APPENDIX B: MATERIALS FROM ITMC

Table 15: Correlations for all participants in the sketching portion of ITMC Study 2. The average correlation is $R=0.34$ ($SD=0.19$). The average was computed using Fisher Z' transformations, as described in Chapter 5.

Large (≥ 0.50)	Medium (≥ 0.30)	Small (≥ 0.20)	Very Small (< 0.20)	Not Correlated
0.65	0.42	0.22	0.13	0.09
0.52	0.32	0.21	0.14	0.08
0.65	0.30	0.20	0.10	
0.47	0.31			
0.53	0.45			
	0.34			
	0.45			
	0.31			
	0.30			
	0.31			
	0.40			

Table 16: Correlations for all participants in the writing portion of ITMC Study 2. The average correlation is $R=0.27$ ($SD=0.16$). The average was computed using Fisher Z' transformations, as described in Chapter 5.

Large (≥ 0.50)	Medium (≥ 0.30)	Small (≥ 0.20)	Very Small (< 0.20)	Not Correlated
0.51	0.38	0.26	0.11	0.02
0.53	0.30	0.24	0.13	0.01
	0.42	0.21	0.12	
	0.30	0.24	0.15	
	0.41	0.12		
	0.30	0.10		
	0.46			
	0.30			
	0.32			
	0.39			