Towards a Reflection in Creative Experience Questionnaire

COREY FORD, Queen Mary University of London, UK
NICK BRYAN-KINNS, Queen Mary University of London, UK

Reflection is underexplored in Creativity Support Tool (CST) research, partly due to its ambiguous nature. We suggest that researchers could benefit from a measure of a CST's capacity to support reflection. To this end, we detail the first stages of development of the Reflection in Creative Experience Questionnaire (RiCE) – a lightweight questionnaire for differentiating between creative user experiences which exhibit more or less moments of reflection. We develop RiCE through i) an expert review of questionnaire items (n=10) and ii) an exploratory factor analysis (n=300) of the reviewed items. We also present a user study testing RiCE (n=58) across two time points (one week apart) with novel interfaces designed for creative writing and music making. Although we do not confirm validity, we identify four factors for RiCE which we suggest are interpretable in a conceptually meaningful way. Our formative studies contribute towards supporting future explorations on reflection with CSTs.

CCS Concepts: • **Applied computing** → **Arts and humanities**; • **General and reference** → *Measurement*; • **Human-centered computing** → *Empirical studies in HCI*; **HCI design and evaluation methods**; *User studies.*

Additional Key Words and Phrases: creative process, creativity, creativity support tools, evaluation, factor analysis, metrics, psychometrics, reflection, reflective practice

ACM Reference Format:
Corey Ford and Nick Bryan-Kinns. 2023. Towards a Reflection in Creative Experience Questionnaire. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3544548.3581077

1 INTRODUCTION

> *"Art is not a reflection of reality, it is the reality of a reflection."* – Jean-Luc Godard, Filmmaker [91, pg. 29]

Designing interfaces to support creativity is an ongoing challenge in Human-Computer Interaction (HCI) research, informing system design across many domains – from supporting children drawing on an iPad to supporting professional artists and designers [31]. Since Fisher [27] and Shneiderman [77] highlighted that there is a need to investigate how computers can support creativity, the HCI sub-field of Creativity Support Tools (CSTs) continues to explore how to design tools supporting aspects of creative user experiences, such as ideation [40, 46]. CST research often overlaps with user experience research, approaching evaluation based on people's subjective experiences. This contrasts more conventional HCI measures of a system's usability which can be inappropriate for creative tasks [45]. For example, conventional HCI measures might consider fast task completion to be a measure of success whereas more time spent on a creative task might be a positive indicator of immersion. For brevity we refer to creative user experiences as creative experiences and distinguish these from non-creative user experiences as tools that support open-ended tasks with no concrete metric of success.

Manuscript submitted to ACM

Despite its importance in the creative process [14, 34, 69, 85, 88], *reflection* is an underexplored aspect of creative experiences. This might be because reflection is ill-defined [6, 28, 61], making it difficult to compare moments of reflection across study conditions or interface prototypes. HCI research has more frequently explored reflection as a desirable aspect in personal informatics [87], slow technology [37] and design processes [76]. We suggest that a measure of reflection in creative experience might be helpful for CST researchers, driving forward investigations in this underexplored area.

In this paper, we detail the first stages of development for the Reflection in Creative Experience questionnaire (RiCE). We aim to design a lightweight self-report tool which can differentiate between creative experiences which exhibit more or less moments of reflection. To this end, we developed an initial item set and reduced this via an expert review (Section 4). We then collected data from 300 people who recently used a creative technology and performed an exploratory factor analysis to reduce the items further, grouping them into factors (Section 5). Next, we conduct a user study to test RiCE with two novel technologies for creative writing and music making (Section 6). To summarise, we offer the following contributions:

- The documentation of the first steps towards a lightweight self-report questionnaire for differentiating between creative experiences which exhibit more or less moments of reflection. We identify four factors for RiCE which we suggest can be interpreted in a conceptually meaningful way [89].
- An exploration of which aspects of reflection might occur in creative tasks, including tasks with interfaces containing aspects of both writing and music making.
- A user study testing RiCE in two HCI contexts related to creative writing and music making, indicating directions for its future development.

2 BACKGROUND

In this section, we discuss literature on reflection and how it could relate to creative contexts (Section 2.1), including discussion on existing measures for reflection (Section 2.1.1). We then introduce Creativity Support Tools (CSTs) (Section 2.2) and techniques for measuring creativity (Section 2.2.1).

2.1 Reflection

There is no consensus on the definition of reflection, partly due to its subjective nature [6, 28, 61]. There is a common understanding of reflection as a thought or consideration [61]. Some interdisciplinary literature suggests reflection has an outcome [3, 11, 12, 81] or is a process applied to clarify uncertain situations [24, 49, 75]. One pragmatic definition is that reflection is "a basic mental process with either a purpose or an outcome or both, that is applied in situations where material is ill-structured or uncertain and where there is no obvious solution" [61, pg. 10]. Norman [64] takes a cognitive perspective, suggesting a generalisation that people experience both moments of *experiential cognition* (expert reactions "without any apparent effort or delay" [64, pg. 23]) and *reflective cognition* (a slower "comparison and contrast, of thought" [64, pg. 26]). Researchers across disciplines have different "conceptual[...] tools for thinking about and analysing reflection" [5, pg. 587]. We suggest below key theories which might be useful in understanding reflection in creative experiences.

Schön [75] developed arguably the most influential theory of reflection used in HCI research [2, 6], introducing *reflection-in-action* (when a person's behaviour does not result in the expected outcome, so they experiment and reflect to solve the issue) and *reflection-on-action* (reflecting after or away from an activity). Slovak, Frauenberger and Fitzpatrick

[79] suggest that Schön's approach might not best support some HCI work as it emphasises the practitioner rather than how to foster a (technology-supported) environment conducive to reflection. Based on case studies in socio-emotional learning, they suggest removing risks from environments (through technology) to facilitate people's reflection processes. Researchers in fields such as education and nursing have developed models of such reflection processes in terms of both how it develops over time [47, 55] and the process of reflecting [3, 11, 12, 24, 49, 81]. An example of the latter is Dewey's [24] model, which views reflection as an inquiry where ideas are formulated, considered, and either accepted or rejected. We suggest these models are helpful as they might indicate how reflection unfolds during creative experiences. For instance, Cho et al. [19] drew upon similar ideas to summarise seven steps for reflection in craft-making – to document, search, observe, organise, compare, connect and iterate.

Designing for reflection became more prominent in HCI around the early 2000s [6], with interest accelerating near 2010, as catalysed by a CHI workshop in 2009 [73] and two review papers [6, 28]. Baumer [5] synthesised interdisciplinary literature on reflection, identifying three dimensions (breakdown, inquiry and transformation) to support discussions on designing for reflection. Fleck and Fitzpatrick [28] also synthesised interdisciplinary literature on reflection to design a pragmatic framework for interaction designers, suggesting how technology could support increasingly sophisticated levels of reflection. Bentvelzen et al. [9] extended Baumer's [6] review in 2022, identifying 98 interactive systems designed to enhance reflection from the ACM digital library (n=52) and the Apple App store (n=46). They identified common design features tied to aspects of reflection such as allowing users to revist their data (to prompt introspection), or to share data to social media (to encourage comparison and conversation).

2.1.1 Measuring Reflection. Measuring reflection is difficult given the lack of a consensus definition [6, 28, 61]. Education and healthcare researchers have developed self-report questionnaires operationalising reflection from different perspectives. A systematic review of 700+ papers [65] recommended the Reflection Questionnaire [44] and Self-Reflection and Insight Scale (SRIS) [32] as most rigorous. The SRIS has informed HCI design considerations for supporting everyday reflection [60], but is not technology focused, instead quantifying people's tendency to self-reflect through three factors: *insight* (people's ability to understand themselves), *engagement in self-reflection* (frequency at which people self-reflect) and *need for reflection* (people's motivation to reflect). It was tested with a confirmatory factor analysis, test-retest study, and a comparison between Psychology students who did and did not keep a diary.

Questionnaires for measuring reflection in HCI contexts are sparse. Although some have been used to examine technology [52, 67, 71], they are not validated nor widely used. Bentvelzen et al. [8] developed the Technology-Supported Reflection Index (TSRI) to quantify levels of reflection afforded by personal informatics systems. Their scale likely provides the measurement closest to our goals in this paper. However, the TSRI is designed for personal informatics, whereas we are interested in assessing people's moments of reflection during a recent creative experience – its questions on (long-term) personal data do not fit our domain of creativity support. Indeed, the TSRI is optimised for interfaces with a functional goal to support people in changing their behaviours given logs of their personal data – at odds with creative interfaces where interaction is open-ended and unpredictable [38]. Items for the TSRI were devised inductively and subjected to an expert review (discussions amongst people knowledgeable in reflection-related HCI). Its factors were then determined through an exploratory factor analysis and examined using two prototypes of a personal informatics dashboard – one designed to prompt more reflection than the other.

2.1.2 Aspects of Reflection. Informed by the literature discussed above, we speculate that the following aspects of reflection might be identified in creative experiences with more or less moments of reflection. We do not claim that we have captured reflection in its entirety, only that our suggestions might help us in exploring reflection in creative contexts.

Indeed, we focus here on research in HCI [5, 9, 28, 64] and on how the process of reflection unfolds [3, 11, 24, 49, 75, 81], and less on the role of reflection in design (e.g. [76]) or knowledge generation (e.g. [35]).

- **Breakdown –** Baumer [5] suggested that reflection occurs in moments of breakdown. Some theories on reflection [24, 49, 75] describe this as where a person's actions map to outcomes against their intuitions.
- **Comparison –** When reflecting, people think back on previous experiences [11, 24, 49, 75, 81] or, as Norman [64] suggests, compare actions to apply in new, uncertain contexts. They might also compare themselves to others [9].
- **Impact –** At the highest level of reflection, Fleck and Fitzpatrick [28] suggest that people consider the broader implications of their actions, including how they influence different people and cultures.
- **Inquiry –** Baumer [5] and Dewey [24] suggest people intentionally generate, test and revise hypothesises iteratively whilst reflecting.
- **Motivation –** For reflection to occur, being given the tools is sometimes not enough. People must also *decide* to engage in reflection [28, 32, 79].
- **Openness –** People remain open to new experiences [49] and paths of inquiry [11, 12] in moments of reflection, acknowledging that variables can change whilst or after reflecting.
- **Transformation –** Many models of the reflection process suggest that people change their understandings [3, 11, 12] and question assumptions when reflecting [5, 49].
- **Trustworthiness –** Norman [64] suggests people sometimes contemplate different information when reflecting. Fleck and Fitzpatrick [28] and Dewey [24] suggest it is the information that is most pragmatic or corroborates with most perspectives that is selected.

2.2 Support for Creativity

Creativity Support Tools (CSTs) – a digital system with features positively influencing people in various stages of the creative process [31] – have been explored in HCI since the early 2000s [31]. Many aspects identified as conducive to creative experiences have been examined to inform CST design [78]. Reflection, although a useful part of the creative process [14, 34, 69, 85, 88], is underexplored in CST research. Some recent examples where CST researchers have discussed how qualities of their tools might support reflection [16, 19, 41, 92] suggest an emerging discourse where a measurement of reflection in creative contexts could be useful. For example, Jonsson and Tholander [41] suggested that the "inconsistent and erroneous" [41, pg. 5] qualities of their code generation tool could be framed as helpful frictions because they encouraged reflection in university students. Emerging sub-genres of CSTs, such as casual creators [20] and its sub-field reflective creators [51], could also be further examples of growing state-of-the-art research areas where reflection is an interesting phenomena [30] and might benefit from ways to measure people's reflection across studies or prototypes.

2.2.1 Measuring Creativity. Attempts to operationalise creativity have roots in Guildford's [33] 1950 address to the American Psychological Association. He suggested that creativity could be measured as the number of divergent uses a participant invents for an "ordinary" object. Critiques of Guildford's approach highlight the context dependent nature of creativity [1, 82]. Such approaches from Psychology also do not always map to CST studies [82]. CST researchers have thus developed their own objective metrics to measure aspects of creativity such as ideation [46] or mutual engagement [13]. Others adopted self-report scales to assess people's feelings of creativity. For example, Wu and Bryan-Kinns [90] used the User Engagement Scale [68] to evaluate their CST's capacity to support non-musicians' engagement in

music making. Recognising the need for a metric of a CST's capacity to support creativity, Cherry and Latulipe [18] developed the Creativity Support Index (CSI). The CSI consists of two parts: i) six eleven-point ordinal item pairs are answered for the creativity-related factors of collaboration, enjoyment, exploration, expressiveness, immersion and results-worth-effort; and ii) fifteen paired comparisons are made across these factors. The total count of factors chosen in the paired comparisons weight the final scores, accounting for which factors are most important in the creative context being assessed. Factors were tested using people's rankings of words related to creativity [17], and further studies supported the CSI's reliability such as a study on people's collaborative use of Google Docs [18] or with artists using drawing software [18].

3 METHOD OVERVIEW

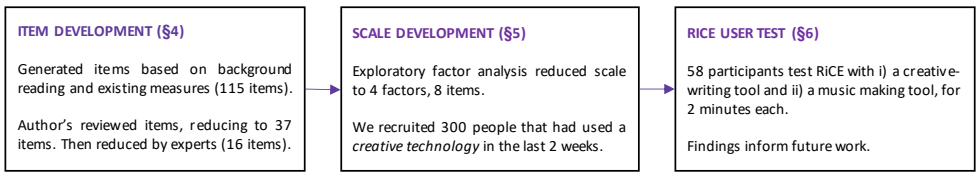


Fig. 1. Overview of the process used to develop and test the first iteration of RiCE.

.

To develop the first iteration of RiCE, we took inspiration from literature on measuring reflection (Section 2.1.1) and measuring creativity (Section 2.2.1). Figure 1 visualises the process for developing RiCE: i) we generated items (statements to be rated by people on an ordinal scale) based on our background reading and reduced these items via reviews by this paper's authors and experts in creativity, ii) we perform an exploratory factor analysis to reduce our items into factors based on 300 people's recent experiences with a creative technology, and iii) we present a user study testing RiCE with two novel interfaces for creative writing and music making to inform future work. Broadly speaking, we follow the approach used to develop the TSRI [8] but apply our analysis to creative tasks inspired by the studies conducted for the CSI [18] because i) as there is no consensus on which aspects of reflection are most valuable in creative contexts we develop our own items and determine factors statistically (as in the TSRI [8]) instead of matching items to factors beforehand (as in the CSI [18]), and ii) we expect RiCE to be used alongside measures such as the CSI [18] which is frequently used to evaluate technologies in creative tasks, and so we explore RiCE in similar contexts. All study phases were approved by the Queen Mary University of London ethics committee for Electronic Engineering and Computer Science. Participants were fully briefed and gave consent. See Appendix for consent forms, questionnaires, data collected and code written in the R[1] programming language for its analysis.

4 ITEM DEVELOPMENT

The first stage of developing RiCE was to determine items that likely indicate moments of reflection in creative experiences. The following subsections detail how we developed our items. We follow a quantitative approach where expert raters (defined in Section 4.2) score items independently. This is relatively quick as multiple experts do not meet to debate nuances as with a qualitative approach, respecting experts' limited time. Section 4.1 details preliminary work used to develop RiCE's initial items, assessed by experts in Section 4.2.

[1]https://www.r-project.org

Table 1. Three examples of items initially generated in the scale development phase. Full list is in the Appendix.

Item	Aspect	Citation	Comment
The system worked in ways which were often puzzling.	Breakdown	[32]	Modified from SRIS to be system oriented.
I identified connections between contrasting ideas and explored this in my creation.	Comparison	[92]	Novel item.
I was able to easily explore other people's ideas.	Openness	[8]	Modified from TSRI to not focus on "data".

4.1 Preliminary Work

To initially develop items, the first author searched through items from existing measures used to evaluate CSTs [18, 39, 68] and measure people's reflection [8, 32, 44, 52, 67, 71] as identified in the literature review. We define an item as a statement to be rated by people against a row of ordinal points. The first author is a male PhD student, exploring how AI might support reflection in music composition. Candidate items were sorted into the aspects of reflection listed in Section 2.1.2, acting as a guide for whether items might indicate moments of reflection. 62 items were rephrased to relate more directly to creativity and reflection, and 49 novel items were written drawing upon the literature above, including recent CST studies discussing reflection (see Section 2.2). In total, 115 items were created. Three examples are shown in Table 1 – a full list is in the Appendix.

To reduce the item set, the first and second author of this paper independently scored each item as "Disagree" (1), "Neutral" (2), or "Agree" (3) against the criterion: "The item appropriately contributes towards assessing if a moment of reflection occurred during a person's creative experience." The second author is a male Professor of Interaction Design in the UK, researching interactive technologies for media and arts. As some items can be interpreted to fit multiple aspects of reflection, the items were shuffled and presented without categorisation – the statistical analysis in Section 5 drives item groupings. A Cronbach's [21] alpha – which is a suitable metric for assessing agreeably between raters when using ordinal data – of .76 was calculated. Following general guidelines [74], we suggest the authors had acceptable agreement.

The authors discussed items where their scoring contrasted. The set was then shortened by removing 60 items where at least 1 author scored "Disagree", excluding 4 items where wording was tweaked. This resulted in 59 items being shuffled and scored again by the authors independently, against a re-worked criterion statement (to be more concrete) of: "The item indicates that a moment of reflection occurred whilst a person was undertaking a creative activity with technology." A Cronbach's [21] alpha of .71 was calculated – we suggest there is acceptable agreement between raters [74]. Of the 59 items, 37 where both authors fully agreed were assessed by 10 experts, as described below.

4.2 Expert Review

We recruited 10 experts through our professional networks. We define experts as people with knowledge of the creative process, where some experience with creativity-related HCI or designing for reflection is desirable. We chose 10 as our sample size because Boateng et al. [10] suggest that typically 5 to 7 expert evaluators are used to develop questionnaires; we round upwards for simplicity. We also tried to represent many creative disciplines to identify items that might be useful to many CST researchers. Table 2 shows the experts' gender (4 Male, 6 Female), age (Mean = 28.2, Med = 28, SD = 4.29), country and summaries of their self-written biographies.

Table 2. Experts' backgrounds who assessed possible RiCE items (see Section 4.2). Biographies are summarised from verbatim biographies found in the Appendix. All participants were instructed to write their biographies to only include information that they consent to be published, as approved by the Queen Mary University of London ethics committee for Electronic Engineering and Computer Science.

ID	Age	Gender	Country	Biography Summarised
P1	29	Female	China	Final year PhD; Musical Interaction; Digital Musical Instrument design; MArch Urban Design; BEng School of Architecture; Teaching experience related to creativity, design and applying technology in these fields.
P2	23	Female	Italy	End of 1st year PhD in AI and Music; Attended conservatoire for piano performance and composition; A-Level Music Technology; Creative Music Technology degree.
P3	34	Female	England	2nd year PhD in Computational Creativity; Examining text-to-image generative AI and Twitter bots; MSc Computer Science; BA(Hons) Fine Art; self-employed (tattoo) artist for several years; ProCreate; Photoshop; Produced paintings for exhibitions.
P4	27	Female	Germany	First year PhD in the Art and Design faculty; Research Assistant in the Computer Science faculty; background in Industrial and Interaction Design; Mentor for first year university students, guiding reflective practices.
P5	33	Female	England	Fourth year PhD; Exploring mindfulness in Interaction Design with AI and Audio.
P6	34	Male	Chile	Third year PhD in Media and Arts Technology; Researching error and music improvisation; experience in web development; Multi-instrumentalist: piano, voice, guitar, venezuelan cuatro; performer & composer.
P7	25	Male	England	Associate Lecturer in Music Technology; BSc(Hons) Music Technology; MSc Creative Technology; Composer of punk and hard rock/metal through to alt-jazz; experience with p5.js and openFrameworks, Unity, Unreal, MaxMSP and Ableton.
P8	29	Female	USA	Fifth year PhD in HCI; Investigating Human-AI Co-Creativity, Ethical AI and Interaction Design; BSc Computer Science and Engineering; Teaching experience in HCI and rapid prototyping.
P9	24	Male	England	Award winning filmmaker; Short films, animation and live action, telling stories on South Asian experiences; Storyboarder; Celtx; Fade-In; Adobe CC Suite (After Effects, Premiere Pro); Davinci Resolve Studio; Final Cut; Clip Studio and TV Paint.
P10	24	Male	Norway	Assistant Film and TV Colourist in a post-production house; VFX turnovers; Grade-matching; Experience working on music videos, short films and TV Series; Baselight; DaVinci Resolve; Premiere Pro.

4.2.1 Procedure. Experts were sent a spreadsheet with the 37 items devised in Section 4.1 and instructions for scoring. Experts were asked to score items "Disagree" (1), "Neutral" (2), or "Agree" (3) against the criterion refined in our preliminary work: "The item indicates that a moment of reflection occurred whilst a person was undertaking a creative activity with technology." A notes column was also provided where experts were encouraged to give further feedback. Items were shuffled for each expert. Experts were reimbursed with a £20 Amazon voucher for their time; we estimate the procedure lasted 30-45 minutes.

4.2.2 Analysis Method. For each item, "Disagree", "Neutral" and "Agree" responses were counted. We list these sorted by the number of "Agree" responses to compare and contrast the highest and lowest scoring items. We also interpret the scoring in the context of the experts' comments. Items for the next phase were retained where more than 7 out of 10 experts selected "Agree". We calculate and interpret inter-rater reliability using Cronbach's [21] alpha as in Section 4.1.

4.2.3 Results. Cronbach's [21] alpha equals .72 – we suggest acceptable agreement between raters [74]. Table 3 lists the highest and lowest scoring items sorted by the number of "Agree" responses – the horizontal line indicates where items are omitted for brevity. Some items with high "Agree" scores relate to iterating (Q23, Q7), self-assessing and selecting actions (Q14, Q11, Q13, Q29). P7 noted that "you can reflect on each interaction to understand why each may not have worked". P10 noted they are "constantly learning and refining techniques". Possibly, a cyclical process of improvement might be important to reflection in creative work. Items regarding worrying about how others perceive your creative work (Q24, Q27, Q33) scored low. P1 suggested that "if the creative activity is about self-expression", worrying about others' perceptions might not indicate reflection. Indeed, P10 did not "mind what others [thought]". Perhaps, moments of reflection in creative activities are personal to creators – some high scoring items relate to personal improvement (Q1, Q19, Q21). Furthermore, experts scored low items on their beliefs being challenged (Q9, Q26). P3 wrote "being challenged != reflecting", whereas P4 suggested such items "better suit reflexivity".

Table 3. The number of experts scoring "Agree", "Netural" or "Disagree" for select items, sorted by the number of "Agree" scores. Items where 7 out of 10 or more experts rated "Agree" were taken forward to the scale development phase. (R) denotes that the item's answer given by a participant in a user study would be reversed.

Q	Item	Total Count		
		"Agree"	"Netural"	"Disagree"
Q23	I often generated, tested and revised ideas.	10	0	0
Q25	Whilst creating, I thought back on some of my past experiences.	10	0	0
Q30	I often reflected on my actions to see whether I could have improved on what I did.	10	0	0
Q7	I found myself iteratively refining and assessing my creative process.	9	1	0
Q14	I pondered over the meaning of what I was doing in relation to my personal experiences.	9	1	0
Q1	I constructively self-assessed my own actions.	9	0	1
Q12	Whilst being creative, it was very interesting to examine different aspect of my creation.	9	0	1
Q5	I sometimes felt doubtful whilst creating my project.	8	2	0
Q11	I made comparisons within the system to consider alternative ways of doing things.	8	2	0
Q13	Whilst being creative, I liked to think about my actions to find alternative ways of doing them.	8	2	0
Q22	I explored my past experiences as a way of understanding new ideas.	8	2	0
Q29	I considered different ways of doing things.	8	1	1
Q2	I considered how my outputs from the system might be interpreted differently in the future	8	1	1
Q35	I often re-examined things I'd already learnt.	7	3	0
Q19	I learned many new things about myself during the experience.	7	2	1
Q21	I often reappraised my experiences with the system so I could learn from them.	7	2	1
Q24	I was not worried about what others may have been thinking about me (R).	3	5	2
Q32	The results of my actions often violated my expectations.	3	3	4
Q27	I didn't really think about how others would perceive my creative process and final product. (R)	3	1	6
Q33	I was not concerned with how others might evaluate my performance (R).	2	5	3
Q26	The system challenged some of my firmly held beliefs.	2	4	4
Q9	Some of my firmly held beliefs were challenged.	1	5	4

5 SCALE DEVELOPMENT

In the previous phase, we shortened a set of 115 items, identifying 16 which might indicate a moment of reflection during a creative experience. Here, we describe an online survey including these 16 items, subjected to an exploratory factor analysis to group these items into factors.

5.1 Participants

Participants were recruited using Prolific[2], an online survey platform. We use Prolific instead of alternatives because it is academic-focused and its participants might show more interest in creativity-related work [66]. We used Prolific's pre-screening features to distribute the survey to participants worldwide who reported to be fluent in English, have a Prolific approval rating above 98%, and use a device with a screen at least weekly. We also required that participants had used a creative technology within the last 2 weeks in our study description. In the study, we offered the creative technologies from Table 1 in Cherry and Latulipe [18, pg. 3] as examples to participants – although, participants could self-report their own creative technology to consider whilst completing our survey also. We continuously recruited until we reached 300 participants after data cleaning (see Section 5.3), recruiting 320 participants in total and rejecting 20. Indeed, Boateng et al. [10] outlined that multiple authors suggest n = 300 as "good" for factor analysis. Participant genders collected in response to the open question "What is your gender?" were: 56.3% Male, 41.3% Female, 1.6% Non-Binary, 0.3% Trans Man and 0.3% None (which we take to mean 'prefer not to say'). Mean age was 29.1 (Med = 26, SD = 9.19). Figure 2 shows the participants' countries – most participants are from Portugal (21.3%), South Africa (18.0%), the UK (12.3%) and Poland (12.3%). Participants were reimbursed an average award of £9.52/hr; it took a mean of 10m 55s to complete the survey (Med = 09m 41s, SD = 5m 24s).

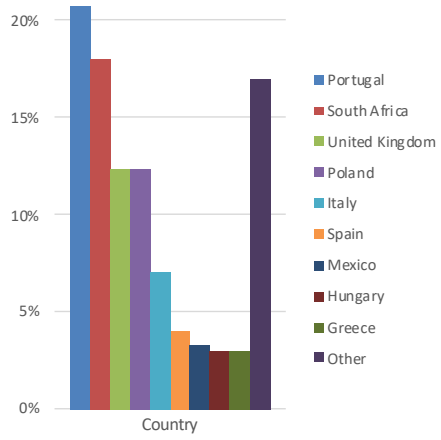


Fig. 2. Participants' countries recruited in the scale development phase.

5.2 Measures & Procedure

To collect measures, we asked the following, in the order listed:

[2]https://www.prolific.co

(1) **Demographics**. As reported above.
(2) **SRIS**. Participants completed the Self-Reflection and Insight Scale (SRIS) scale [32] to evaluate if our sample has a natural tendency to self-reflect (see Section 2.1.1). We calculate 4 means from its factors: *insight*, *engagement in self-reflection*, *need for reflection*, and a total SRIS score.
(3) **Creative Technology**. Participants were asked to select "a creative technology which [they] have used in the last 2 weeks". A drop-down list was provided in the survey based on Table 1 in Cherry and Latulipe [18, pg. 3] but participants could also respond with a free-text description of their own technology. They were then asked to "briefly describe the creative technology... [they selected], how [they] use it, how it supports [their] creativity, and how it supports creativity in general". We used this to clean the data and check the participant's understanding of their chosen technology (see Section 5.3).
(4) **RiCE**. Participants were shown the 16 items identified in our expert review and instructed to rate them "considering their recent experience with their selected creative technology". Each item was placed alongside an 11-point scale with the anchors "Highly Disagree" (0) and "Highly Agree" (10) on either end. We use these anchors to directly mirror the Creativity Support Index (CSI) [18] as it is popular for CST evaluations and thus we might expect RiCE to be used alongside it often. We choose 11-points as multi-point items have been described as easier to use [53] and more points could support test-retest reliability [70].

Finally, participants could offer further comments via an open-ended text box.

5.3 Data Cleaning

We cleaned our data following the advice in [62]. First, we checked participants' understandings of their chosen technology via an opened-ended question (see Section 5.2) – we removed 6 participants who said they had not used a creative technology or did not describe their chosen technology in sufficient detail. Second, we checked for duplicate responses – no responses were identical. Third, we examined a histogram of the survey completion times to identify outliers, removing 6 participants who spent longer than 30 minutes. Fourth, we rejected 8 "flat-liners" [62] who had selected the same option for all items in at least one question block. Respondents were required to complete each question before submission – we had no missing data. This led to our 300 participants (20 out of 320 completed surveys were removed).

5.4 Analysis Method

We report the choice of creative technology and SRIS as descriptive statistics. For the Exploratory Factor Analysis (EFA), we follow Taherdoost, Sahibuddin and Jalaliyoon [83]. Firstly, we assess the sample adequacy by determining whether the Kaiser-Meyer-Olkin (KMO) value is $\geq .7$ [43]. We then assess that Bartlett's [4] test of sphericity is significant ($p < .05$) to indicate that correlations between items are large enough for factor analysis. If these tests are passed, we conduct our EFA with the minimum residual method [54] and oblique rotation because, as with the CSI, we have no reason to believe our items are not correlated [18]. Next, we identify the number of factors where Eigenvalues are > 1.0 as this indicates each factor has a higher variance compared to a single item; we also support this with a scree plot inspection [83]. Then, for each valid factor, we follow Kaiser's [42] rule to select items uniquely correlating with (or loading onto) said factor $\geq .4$. We also calculate Cronbach's [21] alpha to assess inter-item reliability (if items in each factor measure similar constructs), following the guideline that alpha values $\geq .7$ are acceptable, whilst being lenient as scales with few items per construct will naturally yield lower alphas [74] and we aim for RiCE to be lightweight.

5.5 Results

Table 4 shows the creative technologies participants chose when answering our questionnaire. This included software for writing, presentations, photo editing and programming. The SRIS scores are shown in Figure 3 – we interpret these to indicate that participants might be motivated to engage in reflection but do not always understand their insights.

Table 4. Number of participants selecting or suggesting certain creative technologies in the scale development phase.

No. Participants	Creative Technology
20+	MS Word (43); Photoshop (42); Google Docs (29); MS Powerpoint (24)
10+	Visual Studio (15); Adobe Lightroom (15); Blender (13); Adobe Premier Pro (11); AutoCAD (10)
5+	WordPress (8); Google Slides (8); MatLab (7); Illustrator (6); iMovie (6); Paper & Pen (5)
3+	Unity (4) Post-It Notes (3); R Studio (3); Cubase (3)
2	Tableau; Whiteboards; WolframAlpha; Scratch; Final Cut Pro; Adobe After Effects; GarageBand; Prezi; Mendeley; MS Publisher; Cinema 4D; Canva
1	XCode; InkScape; CorelDraw; Logic Pro X; Wikis; MediaWiki; DreamWeaver; Celtx; Obsidian; Clip Studio Paint; Figma; Ableton Live; Arduino; Bear; Kdenlive; Power BI; GIMP; FL Studio; Procreate; TV Paint

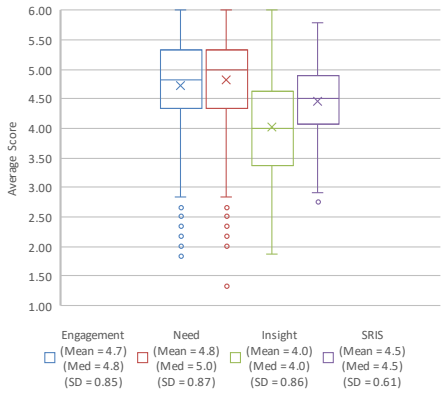


Fig. 3. Overview of the SRIS [32] metrics in the scale development phase.

Table 5. Loadings for the items in the scale development phase. Values > 0.4 are in bold.

Question	Single Factor	Factor 1	Factor 2	Factor 3	Factor 4
Eigenvalue	6.00	2.68	2.05	1.69	2.02
Q11) I made comparisons within the system to consider alternative ways of doing things.	**0.54**	0.00	0.07	0.02	**0.66**
Q23) I often generated, tested and revised ideas.	**0.51**	0.05	-0.16	0.20	**0.57**
Q30) I often reflected on my actions to see whether I could have improved on what I did.	**0.68**	0.38	0.00	0.00	**0.49**
Q12) Whilst being creative, it was very interesting to examine different aspect of my creation.	**0.76**	0.33	0.16	0.12	0.37
Q29) I considered different ways of doing things.	**0.67**	**0.50**	-0.05	0.09	0.28
Q35) I often re-examined things I'd already learnt.	**0.66**	**0.65**	0.08	0.09	-0.02
Q13) Whilst being creative, I liked to think about my actions to find alternative ways of doing them.	**0.76**	**0.88**	0.02	0.03	-0.02
Q7) I found myself iteratively refining and assessing my creative process.	**0.72**	**0.42**	0.12	0.10	0.26
Q1) I constructively self-assessed my own actions.	**0.72**	0.36	0.23	0.08	0.25
Q22) I explored my past experiences as a way of understanding new ideas.	**0.68**	0.12	-0.01	**0.81**	-0.05
Q25) Whilst creating, I thought back on some of my past experiences.	**0.57**	-0.10	0.08	**0.73**	0.06
Q5) I sometimes felt doubtful whilst creating my project.	0.30	0.06	0.20	-0.13	0.29
Q2) I considered how my outputs from the system might be interpreted differently in the future	**0.47**	-0.06	**0.54**	-0.02	0.26
Q14) I pondered over the meaning of what I was doing in relation to my personal experiences.	**0.49**	-0.14	**0.61**	0.20	0.09
Q19) I learned many new things about myself during the experience.	**0.46**	0.08	**0.79**	0.00	-0.13
Q21) I often reappraised my experiences with the system so I could learn from them.	**0.62**	0.20	**0.57**	0.08	0.03

The sampling adequacy was acceptable (KMO = .90) and Bartlett's test of sphericity was significant ($\chi^2(120) = 2110.18$, $p < .000$) – we continue with factor analysis. Table 5 shows the loadings for 4 factors with Eigenvalues > 1 (as supported by our scree plot inspection) explaining 54% of variance. Table 5 also shows our items loading onto a single factor. Factors 1 through 4 explain 17%, 13%, 11% and 13% of variance respectively. As only 2 items loaded onto factor 3 ≥ .4, we selected the top 2 highest loading items from each factor. We also decided to select four factors with two items each because: i) this follows the CSI's [18] format, ii) we aim for RiCE to be as short as possible to minimise participants' fatigue, iii) inspecting the EFA with only 3 factors to increase the number of items per factor identified groupings which we suggest were not easily interpretable [89], and iv) 5 factors did not achieve the necessary Eigenvalues.

Given this, we present the first iteration of RiCE in Table 6, where factors were named based on discussions between this paper's authors. Table 6 also shows the Cronbach's [21] alpha values, suggesting acceptable to moderate inter-item reliability between all factors. We were motivated to retain moderate factors as we only calculated alpha for 2 items making a low value probable [74], the items scored highly in the expert review (see Table 3), and we suggest the factors might be interpreted in a conceptually meaningful way [89].

Table 6. Items and instructions for administering and scoring the first iteration of RiCE. Cronbach's [21] alpha values are also reported giving the inter-item reliability of each factor.

RICE VERSION 1
::: INSTRUCTIONS FOR ADMINISTERING :::
When administering RiCE, each item should be placed along an 11-point scale from "Highly Disagree" (left) to "Highly Agree" (right). Values for each item are zero indexed, i.e., integers from 0 to 10. Please follow the question wording exactly, replacing only the name of your system where indicated. Dimension identifiers (e.g. Cp1), descriptions, and headings should not be visible to participants. Item order should be randomised.
Considering your recent experience of [SYSTEM], please indicate the extent to which you agree with the following statements:
Factor 1 (RiCE-Cp): Reflection on Current Process (α = 0.79) Cp1 (Q13): Whilst being creative, I liked to think about my actions to find alternative ways of doing them. Cp2 (Q35): I often re-examined things I'd already learnt.
Factor 2 (RiCE-Se): Reflection on Self (α=0.68) Se1 (Q19): I learned many new things about myself during the experience. Se2 (Q14): I pondered over the meaning of what I was doing in relation to my personal experiences.
Factor 3 (RiCE-Pa): Reflection on Past Experiences (α=0.77) Pa1 (Q22): I explored my past experiences as a way of understanding new ideas. Pa2 (Q25): Whilst creating, I thought back on some of my past experiences.
Factor 4 (RiCE-Ex): Reflection through Experimentation (α=0.65) Ex1 (Q11): I made comparisons within the system to consider alternative ways of doing things. Ex2 (Q23): I often generated, tested and revised ideas.
All items α = 0.79.
::: INSTRUCTIONS FOR SCORING:::
Following the design of related questionnaires [18, 32, 68], the total RiCE score (out of 10) is calculated as (Cp1+Cp2+Se1+Se2+Pa1+Pa2+Ex1+Ex2) ÷ 8. Each of the 4 factors are calculated as the sum of its items divided by 2 e.g. Reflection on Current Process is (Cp1+Cp2) ÷ 2.

6 RICE USER STUDY

The previous sections detailed the development of the first iteration of RiCE, shown in Table 6. To develop Table 6, items were selected from an Exploratory Factor Analysis (EFA), with factor names derived through discussions between this paper's authors. Here, we conduct a user study to test RiCE in two HCI contexts related to creative writing and music making.

6.1 Participants

We recruited 58 participants through Prolific, with 54 returning to repeat the study procedure 1 week later. We screened for participants who reported to be fluent in English and with an approval rating above 98%. Participants were not required to have previous experience with creative technology as we provide them with novel interfaces (see Section

6.2). Our sample size was based on an apriori calculation in the software G*Power for the Wilcoxon signed-rank test as we collect ordinal data within-subjects (effect size = .5, alpha = .05, power = .95, two-tailed), plus 1 more participant to balance groups. Descriptive statistics for participants' age, gender, compensation and time spent are in Table 7. Figure 4 shows the percentage of participants from each country for both the initial answering of the study and its repetition 1 week later.

Table 7. Descriptive statistics for the participants in the RiCE user study.

Test (n=58)			
Gender		**Male:** 43.1%	**Female:** 56.9%
Compensation			£9.89/hr
	Mean	**Med**	**SD**
Age	27.57	25	8.92
Time Spent	18m 49s	15m 6s	9m 33s
Re-test (n=54)			
Gender		**Male:** 44.1%	**Female:** 55.9%
Compensation			£10.80/hr
	Mean	**Med**	**SD**
Age	27.89	25.5	9.13
Time Spent	16m 55s	15m 5s	7m 35s

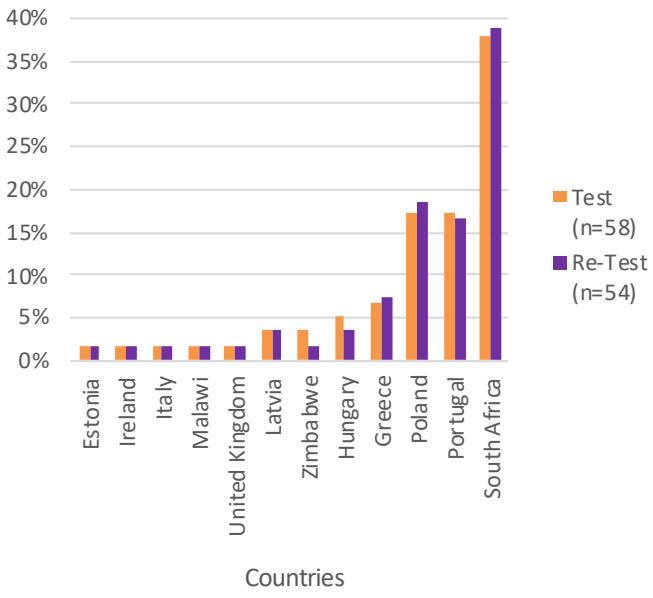


Fig. 4. Participants' countries in the RiCE user study.

6.2 Interfaces

We aspire for RiCE to be used in many creative domains. For this study, we focus on two interfaces we developed to test RiCE in creative experiences containing aspects of writing, music and drawing, representing typical CST activities [18, 31]. We aimed for the interfaces to be simplistic, including the minimal number of features required for people to have a short creative experience. We do not use existing tools as they might require lengthier learning processes and we wanted all participants to have no prior experience with the interfaces. Furthermore, many CST studies focus on evaluating novel high-fidelity prototypes instead of interfaces with a longstanding release [31], making novel interfaces an appropriate subject of formative investigation. Developed with the p5js JavaScript library [59], the interfaces were embedded into the questionnaire alongside descriptions of how to use them, requiring no installation.

6.2.1 Story Sentiment Visualiser. In story-sentiment-visualiser[3], shown in Figure 5a, people are given real-time feedback whilst writing. As text is typed into the interface, each word is allocated a valance score (positive or negative) based on the AFINN-111 data-set [63]. This score is visualised by moving the arrow on the smiley scale at the top of the interface and changing the background colour from red (for negative values) through to green (for positive values). Its design is inspired by principles related to designing for reflection. For example, the visual feedback provides more information than people are usually able to see whilst writing cf. Fleck and Fitzpatrick's [28] design suggestions. Participants using this interface were tasked with writing a positive story (so that their intent is visualised cf. reflective creator design patterns [51]) for two minutes. We explore creative writing as it was used to test the CSI [18] and is an area where reflection is discussed [16, 51]. The task also requires little prior knowledge, making it suitable for novices and likely achievable in a short amount of time.

6.2.2 Sound-sketcher. Sound-sketcher[4], shown in Figure 5b, allows people to draw points which are sonified into a melody, where x-coordinates equal time and y-coordinates equal pitch. People can play and stop the sonification using the play button in the top left corner – their composition is not played in real-time but only when the play button is clicked. They can also switch between a pen and eraser tool, the latter allowing them to remove points. We were inspired by tools used to support novices' music making which similarly turn drawings into sound [22, 25, 56, 84]. As we wanted to validate RiCE for user experiences which include some elements of music and sketching, we thought this style of tool intersected both domains, whilst acknowledging that this is an oversimplification – music and sketching are distinct and broad areas of which sound-sketcher only captures some characteristics. The tool also allows people to create music relatively quickly. Participants were tasked with composing a piece of music for two minutes.

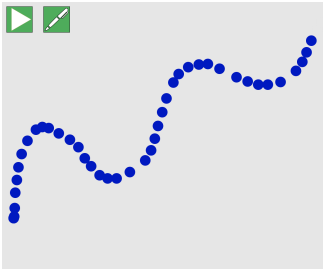
6.3 Measures & Procedure

We asked the following to collect our measures, in the order listed:

(1) **Demographics.** As reported in Section 6.1.
(2) **SRIS**. Participants completed the SRIS [32] – we calculate a total average score.
(3) **Task.** Participants initially use one of our interfaces to complete its associated task. Later the participants use the other interface to complete its associated task. The order is randomised but balanced (50% started with sound-sketcher, 50% with story-sentiment-visualiser). After participants interacted with an interface for 2 minutes they were shown a keyword. Participants had to correctly submit this keyword for payment to be

[3]https://codetta.codes/story-sentiment-visualiser/
[4]https://codetta.codes/sound-sketcher/

(a) A screenshot of Story-Sentiment-Visualiser. Participants were tasked with writing a positive story for two minutes.



(b) A screenshot of Sound-Sketcher. Participants were tasked with writing a music composition for two minutes.

Fig. 5. Screenshots of the novel interfaces used in the RiCE user study.

honoured. This checked that i) participants tested the interface for the required time and ii) that it loaded correctly. No participants were rejected. We do not include training time for the interfaces because i) the tools were designed to be intuitive, and ii) we want to test RiCE with open-ended CSTs where discovery and self-learning is often key [38, 77, 78].

(4) **RiCE.** Participants answered the RiCE items as described in Table 6, considering the interface they had just used. Taking direction from related questionnaires [18, 32, 68], we derive 5 mean averages for *Reflection on Current Process* (RiCE-Cp), *Reflection on Self* (RiCE-Se), *Reflection through Experimentation* (RiCE-Ex), *Reflection on Past Experiences* (RiCE-Pa), and a total RiCE score.

(5) **CSI.** Participants completed the CSI [18] for the interface they had just used to explore how RiCE correlates with the CSI. This included completing both the CSI's item scoring and factor comparison sections (see Section 2.2.1) – we calculate the weighted sum of the means for a total CSI score.

(6) **Repeat.** We repeat steps 3, 4 and 5 for the other interface.

(7) **Comparison.** Participants are asked *"When exploring the 2 interfaces [pictured], with which did you experience the most moments of reflection?"*. This is to test if RiCE or its factors are higher for the interface most participants agree they experienced the most moments of reflection with.

Finally, participants were given an opportunity to offer further comment via an open-ended text box. A week later, we re-invited participants to complete the study procedure again to assess RiCE's test-retest reliability.

6.4 Analysis Method

We describe below the statistical techniques used to test RiCE. Throughout, we assume significance where $p < 0.05$.

6.4.1 Confirmatory Factor Analysis. To test RiCE's factor structure, we ran a Confirmatory Factor Analysis (CFA) on the data collected in the test and re-test conditions for both sound-sketcher and story-sentiment-visualiser. We use the lavaan package for the R programming language [72] (see Appendix) to support reproducibility. Each pair of statements from RiCE were modelled as loading onto their respective factor as identified from our EFA (see Table 6). We used the maximum likelihood estimator with Satorra-Bentler scaling (robust maximum likelihood) as Finney and DiStefano [26] suggested this is appropriate for ordinal data with more than six points.

We examine metrics of our CFA model's fit suggested by Kline [48] which are commonly used and understood across HCI studies such as [15, 23, 57, 86]. These metrics are [10, 48, 58]: a Chi-squared test (to assess the difference between our sample's covariance and the model's covariance), the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) (to assess the ratio between the deviation of our model from the worst fitting model and the deviation of our model from the best fitting model), the Root Mean Squared Error of Approximation (RMSEA) (to measure the degree of our model's misspecification), and the Standardised Root Mean Square Residual (SRMR) (to assess the error between our model's covariance and the sample's covariance). We determine the acceptability of each metric based on suggested criteria: Chi-squared test is not significant ($p \geq 0.05$) [58]; CFI and TLI $\geq .90$ is acceptable [7, 36] and $\geq .95$ is excellent [10, 48]; RMSEA ≤ 0.08 and not significant ($p \geq 0.05$) is acceptable [58]; and SRMR ≤ 0.08 is acceptable [10, 36, 58].

6.4.2 Test-Retest Reliability. Test-retest reliability is the extent to which people's questionnaire responses do not change between points in time. We follow Boateng et al. [10] and the TSRI [8], calculating the Intra-cClass Correlation (ICC) coefficient for RiCE's factors. Points are taken from the first survey responses and 1 week later for both interfaces. We interpret the results following the guidelines in Koo and Mae [50] of poor (ICC $\leq .5$), moderate ($.5 <$ ICC $> .75$), good ($.75 \leq$ ICC $> .9$) and excellent (ICC $\leq .9$).

6.4.3 Differentiation by Known-Groups. Towards understanding how well RiCE captures our intended measure, we examine the difference between RiCE and its factors for the two interfaces. Using the Wilcoxon signed-rank test (as we have ordinal data), we compare the medians for significantly different factors against the count of users who selected the interface they found they had the most moments of reflection with, to determine if the factors move in the same direction.

6.4.4 Comparison with Existing Scales. We identify correlations between RiCE's total score and the total scores of the SRIS [32] and CSI [18] to assess if i) RiCE captures our intended measure and ii) is not simply derivative of these related scales. We assume that higher SRIS scores will occur alongside higher RiCE scores and that higher CSI scores will occur alongside higher RiCE scores. Yet, we expect weak ($\geq .3$ and $< .5$) to moderate ($\geq .5$ and $< .7$) correlations, supporting the notion that, although RiCE is conceptually different, it is still influenced by related factors. Given this, we devised the following hypothesises:

- **H1:** For story-sentiment-visualiser, there will be a weak to moderate positive correlation between RiCE's total score and the SRIS's total score.
- **H2:** For story-sentiment-visualiser, there will be a weak to moderate positive correlation between RiCE's total score and the CSI's total score.
- **H3:** For sound-sketcher, there will be a weak to moderate positive correlation between RiCE's total score and the SRIS's total score.
- **H4:** For sound-sketcher, there will be a weak to moderate positive correlation between RiCE's total score and the CSI's total score.

We only inspect correlations between total scores as opposed to individual factors to i) focus on testing RiCE as a whole and ii) to avoid family-wise type 1 errors on account of multiple tests. We use Spearman's [80] Rho correlation co-efficient as it is suited to ordinal data.

6.5 Results

In this section, we report the results of the statistical tests outlined above.

6.5.1 Confirmatory Factor Analysis. Table 8 shows the fit metrics for the CFA of RiCE. CFI is acceptable in both re-test conditions. SRMR is also acceptable in both re-test conditions and in the test condition for sound-sketcher. There are also some acceptable metrics for the re-test condition of story-sentiment-visualiser for the Chi-squared test and RMSEA. Other metrics do not achieve acceptance.

Table 8. Fit metrics for RiCE's confirmatory factor analysis across conditions and interfaces. Acceptable metrics are in bold.

Timing	Interface	Chi-squared	CFI	TLI	RMSEA	SRMR
	Criterion:	**$p \geq 0.05$**	**≥ 0.9**	**≥ 0.9**	**RMSEA ≤ 0.08; $p \geq 0.05$**	**≤ 0.08**
Test	Sound	$\chi^2(14) = 38.0$, p = 0.00	0.88	0.75	RMESA = 0.17 90% CI [0.11, 0.24] p = 0.00	**0.07**
Re-test	Sound	$\chi^2(14) = 31.0$, p = 0.00	**0.91**	0.82	RMESA = 0.16 90% CI [**0.08**, 0.23] p = 0.01	**0.08**
Test	Story	$\chi^2(14) = 33.3$, p = 0.00	0.89	0.79	RMESA = 0.17 90% CI [0.10, 0.25] p = 0.01	0.09
Re-test	Story	$\chi^2(14) = 20.8$, **p = 0.11**	**0.94**	0.88	RMESA = 0.12, 90% CI [**0.00**, 0.21], **p = 0.15**	**0.07**

6.5.2 Test-Retest Reliability. Table 9 shows the ICCs between the test and re-test measures for RiCE and its factors. For story-sentiment-visualiser, we infer moderate test-retest reliability for all factors, with confidence intervals ranging from poor to moderate, excluding for RiCE-Ex which suggests poor test-retest reliability. Correlations for sound-sketcher also range from poor to moderate. Notably, total RiCE ICCs suggest moderate test-retest reliability for both interfaces.

Table 9. Intra-class correlations between the test and re-test measures for RiCE and its factors. Significant measures in bold.

Interface	RiCE	ICC2	p	CI Lower	CI Upper
Story	RiCE-Ex	.22	.055	.13	.30
Story	**RiCE-Se**	**.52**	**.000**	**.45**	**.59**
Story	**RiCE-Cp**	**.51**	**.000**	**.44**	**.58**
Story	**RiCE-Pa**	**.51**	**.000**	**.43**	**.57**
Story	**RiCE**	**.61**	**.000**	**.55**	**.67**
Sound	**RiCE-Ex**	**.45**	**.000**	**.37**	**.52**
Sound	**RiCE-Se**	**.64**	**.000**	**.58**	**.69**
Sound	**RiCE-Cp**	**.43**	**.000**	**.35**	**.50**
Sound	**RiCE-Pa**	**.47**	**.000**	**.39**	**.54**
Sound	**RiCE**	**.58**	**.000**	**.52**	**.64**

6.5.3 Differentiation by Known-Groups. For participants completing the study for the first time, 60.3% selected that they experienced the most moments of reflection with story-sentiment-visualiser, instead of sound-sketcher (39.7%). This trend continued when participants' completed the study 1 week later (64.8% story-sentiment-visualiser, 35.2% sound-sketcher).

Table 10. Wilcoxon signed-rank tests showing differences across the interfaces, applied for RiCE on both test and re-test. Significant results are in bold.

Timing	RICE	V	p	Median for Story	Median for Sound
Test	**RiCE-Ex**	**501.0**	**.038**	**6.0**	**7.0**
Test	**RiCE-Se**	**974.0**	**.046**	**6.0**	**5.3**
Test	RiCE-Cp	673.5	.891	6.8	7.0
Test	RiCE-Pa	960.5	.111	7.5	6.0
Test	RiCE	828.0	.810	5.9	6.4
Re-test	**RiCE-Ex**	**166.0**	**.000**	**5.8**	**7.5**
Re-test	**RiCE-Se**	**788.5**	**.002**	**6.0**	**4.5**
Re-test	RiCE-Cp	476.0	.483	7.0	6.5
Re-test	RiCE-Pa	685.5	.058	8.0	7.0
Re-test	RiCE	627.0	.693	6.2	6.2

Wilcoxon signed-rank tests were conducted for RiCE and its factors, reported in Table 10. For both the test and re-test responses, RiCE-Ex scores were significantly *lower* for story-sentiment-visualiser than sound-sketcher. Conversely, RiCE-Se scores were significantly *higher* for story-sentiment-visualiser than sound-sketcher.

6.5.4 Comparison with Existing Scales. Here we revisit the hypothesises in Section 6.4.4. For story-sentiment visualiser there is a weak positive correlation between RiCE and the SRIS on test ($r(58) = .36$, $p = .006$) and re-test ($r(54) = .40$, $p = .003$) – we accept H1. There is also a moderate positive correlation between RiCE and the CSI on test ($r(58) = .52$, $p < .000$) and re-test ($r(54) = .66$, $p < 0.000$) – we accept H2. For sound-sketcher, there is a weak positive correlation between the RiCE and SRIS scores on test ($r(58) = .31$, $p = .018$) and re-test ($r(54) = .37$, $p = 0.006$) – we accept H3. Between RiCE and the CSI there is also a moderate positive correlation on test ($r(58) = .54$, $p < 0.000$) and re-test ($r(54) = .67$, $p < 0.000$) – we accept H4.

7 DISCUSSION

To recap, this paper details the initial design of a lightweight questionnaire (RiCE) to differentiate between creative user experiences where people subjectively had more or less moments of reflection. Table 6 shows the first iteration of RiCE designed based on prior literature, an expert review of items and an Exploratory Factor Analysis (EFA). Although we cannot claim validity, the factors found we suggest can be interpreted in a conceptually meaningful way [89]. We also conducted a user study with RiCE, guiding suggestions for future work. Below we discuss our findings, unpacking RiCE's factors in Section 7.1. We also discuss limitations in Section 7.2. Throughout, we consider our work in relation to the literature review (Section 2), and RiCE's factors are referred to using the dimension identifiers in Table 6.

There is some indication that RiCE measures moments of reflection and not a different construct. For instance, RiCE correlated with the Self-Reflection and Insight Scale (SRIS) [32], suggesting that higher RiCE scores occur alongside

more naturally reflective people. In the scale development phase, seven or more experts also fully agreed that the items in RiCE capture reflection. However, our experts' descriptions of their professional background suggest that RiCE's factors might be biased towards music – six out of ten experts worked with music or audio in some form (see Table 2). Nonetheless, the correlation between RiCE and the CSI [18] suggests that reflection occurs more so alongside interfaces which better foster creativity, supporting our assumptions.

The differentiation by known-groups test suggests that RiCE can differentiate between which *types* of reflection people self-report occur more or less frequently when story writing or music making with our novel interfaces – future work is needed to understand if this generalises to other interfaces and tasks. The differences between RiCE-Se's and RiCE-Ex's medians might suggest that moments of *self*-reflection (RiCE-Se) occurred more so with story-sentiment-visualiser, whilst moments of reflecting through experimentation (RiCE-Ex) occur more with sound-sketcher. We speculate that participants scored RiCE-Ex higher for sound-sketcher due to its open-ended interaction – people had to continually evaluate their creations against their own criteria. In contrast, story-sentiment-visualiser offered an evaluation metric through its smiley face slider. This supports Bentvelzen et al.'s [9] suggestion that comparisons to an absolute reference encourage reflection, such as by visualising feedback on people's performance. However, in many creative experiences measures of success are subjective [45]. Perhaps encouraging social comparisons (for example, by sharing work to social media) are thus more useful in creative contexts, supporting reflection through conversations [9]. This said, participants only marginally showed a preference for story-sentiment-visualiser. A study comparing interfaces or interface designs with a stronger split of opinion might show more prominent differences between RiCE and its factors. More work is needed to explore what features distinctly influence reflection in Creativity Support Tools (CST).

RiCE and its factors show moderate to poor test-retest reliability. Given the significant differences between RiCE-Ex and RiCE-Se for story-sentiment-visualiser and sound-sketcher, perhaps test-retest reliability varies between creative disciplines. The Confirmatory Factor Analysis (CFA) fit is better for story-sentiment-visualiser than sound-sketcher, also perhaps indicating that RiCE is task dependent. To improve RiCE across many creative tasks, a similar approach to the Creativity Support Index (CSI) [18] could be tested as discussed in Section 7.2. Furthermore, test-retest reliability might improve if participants investigated our interfaces for longer, or if longer than one week was left between data collection points (to mitigate for learning effects). The stronger fit of our CFA in re-test conditions suggests that RiCE in its current form might more reliably measure reflection when participants are more familiar with a creative interface or task. Indeed, we could speculate that story-sentiment-visualiser and its associated task (writing a story) is possibly more familiar than sound-sketcher's (making music from drawings), hence participants choosing it as most reflective.

7.1 RiCE's Factors

The expert review suggests that moments where people iterate and continually assess their ideas might indicate moments of reflection in creative experiences. This is supported by Dewey [24] and Baumer [5]'s inquiry processes, and many CST researchers [19, 29, 34, 92] who describe how people refine their creative work. Perhaps, Norman [64] and Bentvelzen et al.'s [9] notions that people make comparisons when reflecting is also supported as experts' highly rated items on making comparisons to past experiences (see Table 3, Q25, Q14, and Q22). The EFA suggests people might make comparisons between their personal experiences (RiCE-Se), past experience (RiCE-Pa), and as part of (RiCE-Ex) and looking back on their current process (RiCE-Cp). The distinction between RiCE-Ex and RiCE-Cp could be interpreted as similar to Schön's [75] reflection-in-action (making comparisons between ideas during the creative process, i.e., RiCE-Ex) and reflection-on-action (looking back on one's creative process more broadly, i.e., RiCE-Cp). The

inclusion of the RiCE-Cp factor might also imply that people adapt their creative processes upon reflection, as supported by the transformation stages in some models of reflection [3, 11, 12]. From Slovak, Frauenberger and Fitzpatrick's [79] perspective, RiCE's factors might be too practitioner centred as they do not directly indicate whether aspects of a technology-supported environment encouraged reflection. By design, RiCE instead focuses on one's phenomenological experience but might be applied to compare the effect of different technology-supported environments in future work.

The experts' suggestions from the item development stage and RiCE's self-reflection factor (RiCE-Se) might suggest that creative work is linked to "self expression" (P1). Perhaps, contemplating others' perceptions of one's creative work occurs infrequently, and what Fleck and Fitzpatrick [28] characterised as the highest level of reflection (considering wider impacts), or selecting ideas corroborating with a consensus [24, 28], is less important in creative practices than intuition. This is not to suggest that broader impacts or considering many perspectives is not desirable to encourage in some creative processes, but that they did not seem to occur often during our participants' creative activities. This contrasts the TSRI's [8] finding for personal informatics systems that comparing one's data with an other's data prompts reflection. Maybe, the unimportance of considering others' perspectives can be explained as, when scoring or answering RiCE's items, participants worked alone. It also contrasts the notion to share creative work and encourage social comparisons, discussed above cf. [9].

7.2 Limitations & Future Work

RiCE's reliability is limited to the assessments in our formative user study – we do not claim validity. In particular, the extent to which RiCE's factors are appropriate is limited by our CFA. Our CFA is only indicative of RiCE's fit because i) our scale has the minimum two items per factor [48, pg. 201] whereas three or more items is typically recommended for CFA to avoid specification issues [48], and ii) "the sample size [is relatively speaking] not large" [48, pg. 259]. We tentatively suggest that our current results show potential for future work, tending towards good fit. We also note that our EFA identified factors which, considered with the discussion between RiCE's factors and related work above as well as our expert review, we suggest can be interpreted in a conceptually meaningful way [89]. Given this, further work will explore refining and extending RiCE's current design. We suggest extending RiCE with reversed versions of its current items, increasing the number of items per factor. Indeed, our lower reliability scores may well have been a result of selecting only two items per factor. More items per factor would also allow for refined designs of RiCE to be explored via CFA. For example, correlations between the residual errors of items [48] could guide the design of alternative models for RiCE, later cross-validated. The inclusion of extra-items should be balanced against questionnaire length, however, as RiCE is intended to be used quickly alongside other measures and not increase participants' burden – scales with comparable goals include between 9 and 12 items [8, 18, 39, 68].

Although Prolific supported collecting data across countries, we acknowledge that there are biases in our participants' demographics. RiCE might also be skewed given our participants' biases in technology, and we suggest that the appropriateness of RiCE's items vary across domains. Cherry and Latulipe [18] note how the CSI's comparison questions helped factors generalise across multiple creative domains. Perhaps, designing a comparison section for a future iteration of RiCE would help RiCE's robustness across domains. Extensive studies exploring different demographics and creative domains will also help to support our understandings of reflection within creative experiences. Experiments in specific contexts would enhance RiCE's rigor in these areas and could be paired with qualitative investigations to suggest *why* RiCE produces certain results. Furthermore, we suggest RiCE might be too focused on individual creative activities – further exploration is needed to test RiCE in collaborative work.

8 CONCLUSION

This paper documented the initial development of a lightweight self-report questionnaire for differentiating between creative user experiences which exhibit more or less moments of reflection, named the Reflection in Creative Experience Questionnaire (RiCE). Through an expert review of items and an exploratory factor analysis, we developed the first iteration of RiCE (see Table 6). We identified four factors (reflection on current process, reflection on self, reflection through experimentation, and reflection on past experience) which we suggest can be interpreted in a conceptually meaningful way. We then tested RiCE for tasks with novel interfaces related to creative writing and music making, exploring which aspects of reflection might be useful in these areas. As we cannot claim validity yet, future work will continue developing RiCE and further investigate its properties across creative contexts.

ACKNOWLEDGMENTS

Corey Ford is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported by UK Research and Innovation [grant number EP/S022694/1]. Special thanks to the reviewers for their insightful feedback, Teo Dannemann for reading through an early draft, Simon Colton for his help thinking through concepts related to reflection, and the AI & Music CDTs management team who approved this project's funding.

REFERENCES

[1] Teresa M. Amabile. 1982. Social Psychology of Creativity: A Consensual Assessment Technique. *Journal of Personality and Social Psychology* 43, 5 (1982), 997–1013. https://doi.org/10.1037/0022-3514.43.5.997

[2] Pengcheng An, Saskia Bakker, Sara Ordanovski, Ruurd Taconis, Chris L.E. Paffen, and Berry Eggen. 2019. Unobtrusively Enhancing Reflection-in-Action of Teachers through Spatially Distributed Ambient Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300321

[3] Sue Atkins and Kathy Murphy. 1993. Reflection: A Review of the Literature. *Journal of Advanced Nursing* 18, 8 (1993), 1188–1192. https://doi.org/10.1046/j.1365-2648.1993.18081188.x

[4] M. S. Bartlett. 1950. Tests of Significance in Factor Analysis. *British Journal of Statistical Psychology* 3, 2 (1950), 77–85. https://doi.org/10.1111/j.2044-8317.1950.tb00285.x

[5] Eric P.S. Baumer. 2015. Reflective Informatics: Conceptual Dimensions for Designing Technologies of Reflection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 585–594. https://doi.org/10.1145/2702123.2702234

[6] Eric P.S. Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing Reflection: On the Use of Reflection in Interactive System Design. In *Proceedings of the 2014 Conference on Designing Interactive Systems* (Vancouver, BC, Canada) *(DIS '14)*. Association for Computing Machinery, New York, NY, USA, 93–102. https://doi.org/10.1145/2598510.2598598

[7] Peter M. Bentler and Douglas G. Bonett. 1980. Significance Tests and Goodness of Fit in the Analysis of Covariance Structures. *Psychological Bulletin* 88 (1980), 588–606.

[8] Marit Bentvelzen, Jasmin Niess, Mikołaj P. Woźniak, and Paweł W. Woźniak. 2021. The Development and Validation of the Technology-Supported Reflection Inventory. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 366, 8 pages. https://doi.org/10.1145/3411764.3445673

[9] Marit Bentvelzen, Paweł W. Woźniak, Pia S.F. Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting Reflection in HCI: Four Design Resources for Technologies That Support Reflection. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 2 (mar 2022), 27 pages. https://doi.org/10.1145/3517233

[10] Godfred O Boateng, Torsten B Neilands, Edward A Frongillo, Hugo R Melgar-Quiñonez, and Sera L Young. 2018. Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health* 6 (2018), 149. https://doi.org/10.3389/fpubh.2018.00149

[11] David Boud, Rosemary Keogh, and David Walker. 1985. Promoting Reflection in Learning: A Model. In *Reflection: Turning Experience into Learning*. Nichols Publishing Company, Asbury, USA, 19–40.

[12] Evelyn M Boyd and Ann W Fales. 1983. Reflective Learning: Key to Learning from Experience. *Journal of Humanistic Psychology* 23, 2 (1983), 99–117. https://doi.org/10.1177/002216788323201

[13] Nick Bryan-Kinns and Fraser Hamilton. 2012. Identifying Mutual Engagement. *Behaviour & Information Technology* 31, 2 (2012), 101–125. https://doi.org/10.1080/01449290903377103

[14] Bruce G Buchanan. 2001. Creativity at the Metalevel: AAAI-2000 Presidential Address. *AI Magazine* 22, 3 (2001), 13–13. https://doi.org/10.1609/aimag.v22i3.1569

[15] Xiaowei Cai, Javier Cebollada, and Mónica Cortiñas. 2022. Self-Report Measure of Dispositional Flow Experience in the Video Game Context: Conceptualisation and Scale Development. *International Journal of Human-Computer Studies* 159 (2022), 102746. https://doi.org/10.1016/j.ijhcs.2021.102746

[16] Dashiel Carrera and Sang Won Lee. 2022. Watch Me Write: Exploring the Effects of Revealing Creative Writing Process through Writing Replay. In *Proceedings of the Fourteenth ACM Conference on Creativity and Cognition* (Venice, Italy) *(C&C '22)*. Association for Computing Machinery, New York, NY, USA, 146–160. https://doi.org/10.1145/3527927.3532806

[17] Erin A. Carroll, Celine Latulipe, Richard Fung, and Michael Terry. 2009. Creativity Factor Evaluation: Towards a Standardized Survey Metric for Creativity Support. In *Proceedings of the Seventh ACM Conference on Creativity and Cognition* (Berkeley, California, USA) *(C&C '09)*. Association for Computing Machinery, New York, NY, USA, 127–136. https://doi.org/10.1145/1640233.1640255

[18] Erin Cherry and Celine Latulipe. 2014. Quantifying the Creativity Support of Digital Tools through the Creativity Support Index. *ACM Transactions on Computer-Human Interaction* 21, 4, Article 21 (Jun 2014), 25 pages. https://doi.org/10.1145/2617588

[19] Yen-Ting Cho, Yen-Ling Kuo, Yen-Ting Yeh, Huai-Hsuan Liang, and Yu-Ting Li. 2022. Motion-Centric Tools to Reflect on Digital Creative Experiences and Created Outputs. In *Proceedings of the Fourteenth ACM Conference on Creativity and Cognition* (Venice, Italy) *(C&C '22)*. Association for Computing Machinery, New York, NY, USA, 234–246. https://doi.org/10.1145/3527927.3531454

[20] Kate Compton and Michael Mateas. 2015. Casual Creators. In *Proceedings of the 6th International Conference on Computational Creativity* (Park City, Utah, USA). Brigham Young University, Provo, Utah, 228–235. https://computationalcreativity.net/iccc2015/proceedings/10_2Compton.pdf

[21] Lee J Cronbach. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16, 3 (1951), 297–334. https://doi.org/10.1007/BF02310555

[22] Teodoro Dannemann and Mathieu Barthet. 2021. SonicDraw: A Web-based Tool for Sketching Sounds and Drawings. In *Proceedings of the International Computer Music Conference 2021* (Santiago, Chile). International Computer Music Association Inc., San Francisco, CA, USA, 301–308.

[23] Nicholas David Bowman, JihHsuan Tammy Lin, and Chieh Wu. 2021. A Chinese-Language Validation of the Video Game Demand Scale (VGDS-C): Measuring the Cognitive, Emotional, Physical, and Social Demands of Video Games. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 117, 10 pages. https://doi.org/10.1145/3411764.3445348

[24] John Dewey. 1933. *How We Think*. Prometheus Books, Buffalo, New York.

[25] Morwaread M Farbood, Egon Pasztor, and Kevin Jennings. 2004. Hyperscore: A Graphical Sketchpad for Novice Composers. *IEEE Computer Graphics and Applications* 24, 1 (2004), 50–54. https://doi.org/10.1109/MCG.2004.1255809

[26] Sara J Finney and Christine DiStefano. 2006. Non-normal and Categorical Data in Structural Equation Modeling. *Structural Equation Modeling: A Second Course* 10, 6 (2006), 269–314.

[27] Gerhard Fischer. 2004. Social Creativity: Turning Barriers into Opportunities for Collaborative Design. In *Proceedings of the Eighth Conference on Participatory Design: Artful Integration: Interweaving Media, Materials and Practices - Volume 1* (Toronto, Ontario, Canada) *(PDC 04)*. Association for Computing Machinery, New York, NY, USA, 152–161. https://doi.org/10.1145/1011870.1011889

[28] Rowanne Fleck and Geraldine Fitzpatrick. 2010. Reflecting on Reflection: Framing a Design Landscape. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction* (Brisbane, Australia) *(OZCHI '10)*. Association for Computing Machinery, New York, NY, USA, 216–223. https://doi.org/10.1145/1952222.1952269

[29] Corey Ford and Nick Bryan-Kinns. 2022. Identifying Engagement in Children's Interaction Whilst Composing Digital Music at Home. In *Proceedings of the Fourteenth ACM Conference on Creativity and Cognition* (Venice, Italy) *(C&C '22)*. Association for Computing Machinery, New York, NY, USA, 443–456. https://doi.org/10.1145/3527927.3532794

[30] Corey Ford and Nick Bryan-Kinns. 2022. Speculating on Reflection and People's Music Co-Creation with AI. In *Workshop on Generative AI and HCI at the CHI Conference on Human Factors in Computing Systems 2022*. https://qmro.qmul.ac.uk/xmlui/handle/123456789/80144

[31] Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, Michael Mose Biskjaer, and Peter Dalsgaard. 2019. Mapping the Landscape of Creativity Support Tools in HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3290605.3300619

[32] A. M. Grant, J. Franklin, and P. Langford. 2002. The Self-Reflection and Insight Scale: A New Measure of Private Self-consciousness. *Social Behavior and Personality: An International Journal* 30, 8 (2002), 821–836. https://doi.org/10.2224/sbp.2002.30.8.821

[33] J Guilford. 1950. Creativity. *American Psychologist* 5, 9 (1950), 444–454.

[34] Christina Guillaumier. 2016. Reflection as Creative Process: Perspectives, Challenges and Practice. *Arts and Humanities in Higher Education* 15, 3-4 (2016), 353–363. https://doi.org/10.1177/1474022216647381

[35] Jürgen Habermas. 1987. *Knowledge and Human Interests*. Polity Press, Cambridge, UK.

[36] Joseph F. Hair, Rolph E. Anderson, Ronald L. Tatham, and William C. Black. 1995. *Multivariate Data Analysis (4th Ed.): With Readings*. Prentice-Hall, Inc., USA.

[37] Lars Hallnäs and Johan Redström. 2001. Slow Technology – Designing for Reflection. *Personal and Ubiquitous Computing* 5, 3 (Jan 2001), 201–212. https://doi.org/10.1007/PL00000019

[38] Thomas T. Hewett. 2005. Informing the Design of Computer-based Environments to Support Creativity. *International Journal of Human-Computer Studies* 63, 4 (2005), 383–409. https://doi.org/10.1016/j.ijhcs.2005.04.004

[39] Susan A Jackson, Andrew J Martin, and Robert C Eklund. 2008. Long and Short Measures of Flow: The Construct Validity of the FSS-2, DFS-2, and New Brief Counterparts. *Journal of Sport and Exercise Psychology* 30, 5 (2008), 561–587. https://doi.org/10.1123/jsep.30.5.561

[40] Youngseung Jeon, Seungwan Jin, Patrick C. Shih, and Kyungsik Han. 2021. FashionQ: An AI-Driven Creativity Support Tool for Facilitating Ideation in Fashion Design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 576, 18 pages. https://doi.org/10.1145/3411764.3445093

[41] Martin Jonsson and Jakob Tholander. 2022. Cracking the Code: Co-Coding with AI in Creative Programming Education. In *Creativity and Cognition* (Venice, Italy) *(C&C '22)*. Association for Computing Machinery, New York, NY, USA, 5–14. https://doi.org/10.1145/3527927.3532801

[42] Henry F. Kaiser. 1960. The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement* 20, 1 (1960), 141–151. https://doi.org/10.1177/001316446002000116

[43] Henry F Kaiser. 1970. A Second Generation Little Jiffy. *Psychometrika* 35, 4 (1970), 401–415. https://doi.org/10.1007/BF02291817

[44] David Kember, Doris Y. P. Leung, Alice Jones, Alice Yuen Loke, Jan McKay, Kit Sinclair, Harrison Tse, Celia Webb, Frances Kam Yuet Wong, Marian Wong, and Ella Yeung. 2000. Development of a Questionnaire to Measure the Level of Reflective Thinking. *Assessment & Evaluation in Higher Education* 25, 4 (2000), 381–395. https://doi.org/10.1080/713611442

[45] Andruid Kerne, Andrew M. Webb, Celine Latulipe, Erin Carroll, Steven M. Drucker, Linda Candy, and Kristina Höök. 2013. Evaluation Methods for Creativity Support Environments. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 3295–3298. https://doi.org/10.1145/2468356.2479670

[46] Andruid Kerne, Andrew M. Webb, Steven M. Smith, Rhema Linder, Nic Lupfer, Yin Qu, Jon Moeller, and Sashikanth Damaraju. 2014. Using Metrics of Curation to Evaluate Information-Based Ideation. *ACM Transactions on Computer-Human Interaction* 21, 3, Article 14 (Jun 2014), 48 pages. https://doi.org/10.1145/2591677

[47] Patricia M King and Karen Strohm Kitchener. 1994. Developing Reflective Judgment: Understanding and Promoting Intellectual Growth and Critical Thinking in Adolescents and Adults. In *Jossey-Bass Higher and Adult Education Series and Jossey-Bass Social and Behavioral Science Series*. ERIC Institute of Educational Sciences, Sansome Street, San Francisco.

[48] Rex B Kline. 2015. *Principles and Practice of Structural Equation Modeling*. Guilford Publications, New York, USA.

[49] David A Kolb. 1984. *Experiential Learning: Experience as the Source of Learning and Development*. Englewood Cliffs, New Jersey: Prentice-Hall.

[50] Terry K. Koo and Mae Y. Li. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine* 15, 2 (2016), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

[51] Max Kreminski and Michael Mateas. 2021. Reflective Creators. In *Proceedings of the Twelfth International Conference on Computational Creativity (ICCC 2021)*. Mexico City. https://mkremins.github.io/publications/ReflectiveCreators_ICCC2021.pdf

[52] Theodora C Levine. 2014. The Use of Blogging in Tertiary Healthcare Educational Settings to Enhance Reflective Learning in Nursing Leadership. *Journal for Nurses in Professional Development* 30, 6 (2014). https://doi.org/10.1097/NND.0000000000000103

[53] James R. Lewis and Oundefineduzhan Erdinç. 2017. User Experience Rating Scales with 7, 11, or 101 Points: Does it Matter? *Journal of Usability Studies* 12, 2 (Feb 2017), 73–91.

[54] Susana Lloret-Segura, Adoracion Ferreres-Traver, Ana Hernandez-Baeza, and Ines Tomas-Marco. 2014. Exploratory item factor analysis: A practical guide revised and updated. *Anales de Psicología* 30, 3 (2014), 1151–1169.

[55] Patrick G. Love and Victoria L. Guthrie. 1999. King and Kitchener's Reflective Judgment Model. *New Directions for Student Services* 1999, 88 (1999), 41–51. https://doi.org/10.1002/ss.8804

[56] Sebastian Löbbers, Mathieu Barthet, and György Fazekas. 2021. Sketching Sounds: An Exploratory Study on Sound-shape Associations. In *Proceedings of the International Computer Music Conference 2021* (Santiago, Chile). International Computer Music Association Inc., San Francisco, CA, USA, 275–280.

[57] Guido Makransky, Lau Lilleholt, and Anders Aaby. 2017. Development and Validation of the Multimodal Presence Scale for Virtual Reality Environments: A Confirmatory Factor Analysis and Item Response Theory Approach. *Computers in Human Behavior* 72 (2017), 276–285. https://doi.org/10.1016/j.chb.2017.02.066

[58] Masaki Matsunaga. 2010. How to factor-analyze your data right: do's, don'ts, and how-to's. *International Journal of Psychological Research* 3, 1 (Jun. 2010), 97–110. https://doi.org/10.21500/20112084.854

[59] Lauren McCarthy, Casey Reas, and Ben Fry. 2015. *Getting Started with P5.js: Making Interactive Graphics in JavaScript and Processing*. Maker Media Inc., San Francisco, CA, USA.

[60] Ine Mols, Elise van den Hoven, and Berry Eggen. 2016. Informing Design for Reflection: An Overview of Current Everyday Practices. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (Gothenburg, Sweden) *(NordiCHI '16)*. Association for Computing Machinery, New York, NY, USA, Article 21, 10 pages. https://doi.org/10.1145/2971485.2971494

[61] Jennifer A Moon. 2013. *Reflection in Learning and Professional Development: Theory and Practice*. Routledge, New York, USA.

[62] Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge. 2014. *Survey Research in HCI*. Springer New York, New York, NY, 229–266. https://doi.org/10.1007/978-1-4939-0378-8_10

[63] Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. In *Workshop on Making Sense of Microposts: Big Things Come in Small Packages* (Heraklion, Crete, Greece). 93–98. http://arxiv.org/abs/1103.2903

[64] Donald A Norman. 1993. *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine.* Addison-Wesley Publishing Company, Reading, Massachusetts.

[65] Su Min Ooi, Paul Fisher, and Siân Coker. 2021. A Systematic Review of Reflective Practice Questionnaires and Scales for Healthcare Professionals: A Narrative Synthesis. *Reflective Practice* 22, 1 (2021), 1–15. https://doi.org/10.1080/14623943.2020.1801406

[66] Jonas Oppenlaender, Kristy Milland, Aku Visuri, Panos Ipeirotis, and Simo Hosio. 2020. Creativity on Paid Crowdsourcing Platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376677

[67] Sharleen L. O'Reilly and Julia Milner. 2015. Transitions in Reflective Practice: Exploring Student Development and Preferred Methods of Engagement. *Nutrition & Dietetics* 72, 2 (2015), 150–155. https://doi.org/10.1111/1747-0080.12134

[68] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and New UES Short Form. *International Journal of Human-Computer Studies* 112 (2018), 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004

[69] Ulla Pohjannoro. 2016. Capitalising on Intuition and Reflection: Making Sense of a Composer's Creative Process. *Musicae Scientiae* 20, 2 (2016), 207–234. https://doi.org/10.1177/1029864915625727

[70] Carolyn C Preston and Andrew M Colman. 2000. Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences. *Acta Psychologica* 104, 1 (2000), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

[71] Bettina Renner, Joachim Kimmerle, Dominik Cavael, Volker Ziegler, Lisa Reinmann, and Ulrike Cress. 2014. Web-Based Apps for Reflection: A Longitudinal Study With Hospital Staff. *Journal of Medical Internet Research* 16, 3 (17 Mar 2014), 85. https://doi.org/10.2196/jmir.3040

[72] Yves Rosseel. 2012. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* 48, 2 (2012), 1–36. https://doi.org/10.18637/jss.v048.i02

[73] Corina Sas and Alan Dix. 2009. Designing for Reflection on Experience. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI EA '09)*. Association for Computing Machinery, New York, NY, USA, 4741–4744. https://doi.org/10.1145/1520340.1520730

[74] Martin Schrepp. 2020. On the Usage of Cronbach's Alpha to Measure Reliability of UX Scales. *Journal of Usability Studies* 15, 4 (2020).

[75] Donald A. Schön. 1983. *The Reflective Practitioner: How Professionals Think in Action.* Basic Books inc., London.

[76] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph 'Jofish' Kaye. 2005. Reflective Design. In *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility* (Aarhus, Denmark) *(CC '05)*. Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/1094562.1094569

[77] Ben Shneiderman. 2002. Creativity Support Tools. *Communications of the ACM* 45, 10 (Oct 2002), 116–120. https://doi.org/10.1145/570907.570945

[78] Ben Shneiderman, Gerhard Fischer, Mary Czerwinski, Mitch Resnick, Brad Myers, Linda Candy, Ernest Edmonds, Mike Eisenberg, Elisa Giaccardi, Thomas T. Hewett, Pamela Jennings, Bill Kules, Kumiyo Nakakoji, Jay Nunamaker, Randy Pausch, Ted Selker, Elisabeth Sylvan, and Michael Terry. 2006. Creativity Support Tools: Report From a U.S. National Science Foundation Sponsored Workshop. *International Journal of Human-Computer Interaction* 20, 2 (2006), 61–77. https://doi.org/10.1207/s15327590ijhc2002_1

[79] Petr Slovák, Christopher Frauenberger, and Geraldine Fitzpatrick. 2017. Reflective Practicum: A Framework of Sensitising Concepts to Design for Transformative Reflection. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2696–2707. https://doi.org/10.1145/3025453.3025516

[80] C. Spearman. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* 15, 1 (1904), 72–101. http://www.jstor.org/stable/1412159

[81] Norman Steinaker and M Robert Bell. 1975. A Proposed Taxonomy of Educational Objectives: The Experiential Domain. *Educational Technology* 15, 1 (1975), 14–16.

[82] R Sternberg and Todd Lubart. 1999. The Concept of Creativity: Prospects and Paradigms. In *Handbook of Creativity*, R Sternberg (Ed.). Cambridge University Press, New York, Chapter 1, 3–15.

[83] Hamed Taherdoost, Shamsul Sahibuddin, and Neda Jalaliyoon. 2014. Exploratory Factor Analysis; Concepts and Theory. In *Advances in Applied and Pure Mathematics*, Jerzy Balicki (Ed.). Mathematics and Computers in Science and Engineering Series, Vol. 27. WSEAS, 375–382. https://hal.archives-ouvertes.fr/hal-02557344

[84] Jean-Baptiste Thiebaut, Patrick G. T. Healey, and Nick Bryan Kinns. 2008. Drawing Electroacoustic Music. In *Proceedings of the International Computer Music Conference 2008* (Belfast, Ireland). International Computer Music Association Inc., San Francisco, CA, USA, 7 pages.

[85] Shelley Tracey. 2007. Creative Reflection, Creative Practice: Expressing the Inexpressible. In *Proceedings of the 2007 Conference on Creativity or Conformity* (Cardiff, Wales, UK). University of Wales, 8–10.

[86] Jukka Vahlo and Veli-Matti Karhulahti. 2020. Challenge Types in Gaming Validation of Video Game Challenge Inventory (CHA). *International Journal of Human-Computer Studies* 143 (2020), 102473. https://doi.org/10.1016/j.ijhcs.2020.102473

[87] Elisabeth T. Kersten van Dijk, Joyce H.D.M. Westerink, Femke Beute, and Wijnand A. IJsselsteijn. 2017. Personal Informatics, Self-Insight, and Behavior Change: A Critical Review of Current Literature. *Human–Computer Interaction* 32, 5-6 (2017), 268–296. https://doi.org/10.1080/07370024.2016.1276456

[88] Geraint A Wiggins. 2006. Searching for Computational Creativity. *New Generation Computing* 24, 3 (2006), 209–222. https://doi.org/10.1007/BF03037332

[89] Roger L Worthington and Tiffany A Whittaker. 2006. Scale Development Research: A Content Analysis and Recommendations for Best Practices. *The Counseling Psychologist* 34, 6 (2006), 806–838.

[90] Yongmeng Wu and Nick Bryan-Kinns. 2017. Supporting Non-Musicians? Creative Engagement with Musical Interfaces. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition* (Singapore) *(C&C '17)*. Association for Computing Machinery, New York, NY, USA, 275–286. https://doi.org/10.1145/3059454.3059457

[91] Gene Youngblood. 1998. Jean-Luc Godard: No Difference between Life and Cinema. In *Jean-Luc Godard: Interviews*, David Sterritt (Ed.). University Press of Mississippi, Chapter 5, 9–49.

[92] Paulina Yurman. 2021. Fluid Speculations: Drawing Artefacts in Watercolour as Experimentation in Research Through Design. In *Proceedings of the Thirteenth ACM Conference on Creativity and Cognition* (Virtual Event) *(C&C '21)*. Association for Computing Machinery, New York, NY, USA, Article 38, 13 pages. https://doi.org/10.1145/3450741.3466777

A APPENDIX

All Appendix material can downloaded from the ACM Digital Library or found at: https://github.com/thecoreyford/Towards-RiCE.