The Development and Validation of the Technology-Supported Reflection Inventory

Marit Bentvelzen
Utrecht University
Utrecht, the Netherlands
m.bentvelzen@uu.nl

Mikołaj P. Woźniak
Lodz University of Technology
Lodz, Poland
mpwozniak@ubicomp.pl

Jasmin Niess
University of Bremen
Bremen, Germany
niessj@uni-bremen.de

Paweł W. Woźniak
Utrecht University
Utrecht, the Netherlands
p.w.wozniak@uu.nl

ABSTRACT

Reflection is an often addressed design goal in Human-Computer Interaction (HCI) research. An increasing number of artefacts for reflection have been developed in recent years. However, evaluating if and how an interactive technology helps a user reflect is still complex. This makes it difficult to compare artefacts (or prototypes) for reflection, impeding future design efforts. To address this issue, we developed the *Technology-Supported Reflection Inventory* (TSRI), which is a scale that evaluates how effectively a system supports reflection. We first created a list of possible scale items based on past work in defining reflection. The items were then reviewed by experts. Next, we performed exploratory factor analysis to reduce the scale to its final length of nine items. Subsequently, we confirmed test-retest validity of our instrument, as well as its construct validity. The TSRI enables researchers and practitioners to compare prototypes designed to support reflection.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; **HCI design and evaluation methods**.

KEYWORDS

personal informatics, reflection, tracking, construal

1 INTRODUCTION

As an increasing number of systems around us aim to increase our well-being, reflection is a concept that gains more and more relevance. While reflection is complex and can be interpreted in a variety of ways, it is regarded as beneficial and desired [6]. As a consequence, recent years have seen an ever-increasing interest in HCI to design technology that supports reflection in users. These design efforts span a wide spectrum of applications, including mental health [27], personal informatics [15], health [3, 4], as well as reflection on daily life [19]. There is an emerging challenge in the field to understand what particular design qualities can foster reflection.

However, the evaluation of reflection support technologies is often limited. Consequently, it is often unclear how a certain design affected the user's ability to reflect. A systematic review by Baumer et al. [5] showed that there was no consensus in the field regarding the evaluation of reflection support technology. Studies that used quantitative assessments rarely measured reflection, but rather focused on user experience. Qualitative evaluation is an often-used alternative. Yet, this approach, encounters similar difficulties, namely, the lack of a clear definition of reflection and not asking users directly about reflection. The lack of proper evaluation methods for reflection support technology makes it difficult to compare different design alternatives. This, in turn, complicates future design efforts. Consequently, there is a need for building rapid and standardised means of evaluating whether a system can support reflection.

To this end, this paper reports on the *Technology-Supported Reflection Inventory* (TSRI), a scale that evaluates how an interactive system fosters personal, data-driven reflection. We first investigated past work in reflection theory, existing reflection scales and studies regarding reflection in personal informatics. Subsequently, we generated initial items for the scale, which were then subjected to two rounds of expert reviews. We then performed exploratory factor analysis in order to reduce the number of items and obtain the final scale. Finally, we evaluated the TSRI by testing its reliability and construct validity. Our work offers the first, to our knowledge, validated system-centric scale for evaluation of HCI technologies designed for reflection. We contribute the scale for its future use by the research and practitioner community along with a systematic validation of the instrument.

This work is organised as follows, we begin by discussing related work associated with designing for reflection in HCI and existing scales for reflection. Next, we describe the process of developing the scale and the use of an exploratory factor analysis to reduce and optimise the number of scale items. Subsequently, we report on the validation of the scale, and, finally, discuss on the use and limitations of the TSRI.

2 RELATED WORK

In this section, we first provide an overview of the general understanding of reflection within HCI. We then review past work in the area of personal informatics and showcase that there is a need for instruments that enable researches to evaluate technologies designed for reflection.

2.1 Understanding Reflection

Reflection is a recurring research theme in HCI research. Yet, as a systematic review by Baumer et al. [5] showed, few papers (30 out of 76) clearly define the concept of reflection. Consequently, within the HCI community, there are multiple understandings of reflection in place [6]. Several conceptual and theoretical accounts of reflection co-exist in HCI [5, 25].

The majority of HCI papers employ Schön's [24] notion of reflection-in-action and reflection-on-action. However, there is no consensus on the definition of reflection [5], and past research also used alternative definitions by Dewey [6, 8], Moon [10], Mezirow [18, 22, 30] and Boud [23]. These four definitions regard reflection as a process that happens in hindsight, after the experience has taken place. These definitions are more in line with reflection-on-action, as opposed to Schön's notion of reflection-in-action. This multitude of approaches used shows how the field desires to conceptualise reflection in order to understand how to design technologies that could support users in reflecting. This work contributes to building a better understanding of reflection in the context of interactive technologies by building a scale that identifies reflection-supporting qualities in designs.

2.2 Designing for Reflection

Reflection is also a concept in past studies of of personal informatics experiences. Li's *Stage-Based Model of Personal Informatics Systems* [15] considers reflection to be a stage that occurs after the *preparation*, *collection* and *integration* of personal data. Subsequently, after reflecting on personal data a user then moves on to *action*, the final stage of the stage-based model [15]. Li's model was later extended by Epstein et al., who proposed the *Lived Informatics Model of Personal Informatics* [9]. This model includes additional stages, such as deciding to track, the selection of tools, and defines *tracking and acting* as an ongoing process of collection, integration and tracking [9]. These two models offer a high-level overview, but do not provide an in-depth understanding of what the reflection stage entails. As previously noted by Baumer [5], these models carry an implicit assumption that reflection automatically occurs as long as a user has access to data that has been 'prepared, combined, and transformed' [5]. However, this conflicts with reflection theories, that highlight that reflection often does not occur automatically, but needs to be encouraged, as observed by Slovak [25]. This is also

in line with Niess and Woźniak [21], who concluded that reflection with technology support was instrumental for achieving fitness tracker goals.

Concurrently, the HCI field contributed a number of systems that were reported to effectively facilitate reflection, e.g. [1, 3, 4, 13]. An example of such a system is *Trackly*, designed by Ayobi et al. [3]. Trackly is a mobile app that helps patients with Multiple Sclerosis to manually track their symptoms and to reflect on the collected data through visualisations. Whereas Trackly is mostly used in an individualistic setting, there are also systems that use social interaction to enhance reflection, such as *MoodMap*, designed by Rivera-Pelayo et al. [23]. MoodMap is a computer application that lets users track and reflect on their mood at work. In this sense, MoodMap is comparable to Trackly, yet, this system differs in that it also lets users compare their mood to that of colleagues. Another example is *Reveal-it!* designed by Valkanova et al. [29], which is an interactive public display that encourages participants to reflect on their energy consumption to increase social awareness and discourse. The system lets a user voluntarily enter their energy consumption data, which is then visualised as a sunburst representation on the public display. This makes it possible to compare one's energy consumption with others as well as more general statistics for neighbourhoods, which in turn can lead to reflection.

The examples above suggests that there are certain qualities to interactive technologies that determine how effectively they can support reflection. Reflection systems address different application areas and use different interaction techniques. Despite this diversity, reflection appears to a unifying quality. Further, the design process of such technologies would benefit from comparing prototypes at different stages of the design to determine optimal alternatives. However, as Baumer [5] remarked, there are no established ways of evaluating systems that support reflection. Our work aims to contribute to a better understanding of reflection by building a validated evaluation method for systems that support reflection.

2.3 Measuring Reflection

As we aim to develop a measure of reflection useful for HCI work, we first examine measures of reflection in other fields. Several scales were developed with the intention to measure a persons reflective capacity. The Self-Reflection and Insight Scale (SRIS) [11] consists of three factors: *engagement in self-reflection*, *need for reflection* and *insight*. The scale was developed to provide researchers with an instrument to better understand of the socio-cognitive and meta-cognitive processes central to individual change [11].

Further, the Groningen Reflection Ability Scale (GRAS) [2] was developed to measure the personal reflection ability of medical students. The GRAS consists of 23 items, such as: *'I can see an experience from different standpoints'* and *'I take a closer look at my own habits of thinking'*. The GRAS measures a persons ability to reflect on three dimensions: self-reflection, empathetic reflection and reflective communication. Self-reflection is the introspective aspect of reflection, in which a person frames one's feelings, thoughts, beliefs and norms. In contrast, empathetic reflection is the social extension of self-reflection, the kind of reflection in which a person considers the position of others. Additionally, the GRAS measures

reflective communication, which is the behavioural expression of both self-reflection and empathetic reflection [2].

The SRIS and GRAS both measure a person's reflective capacity. They are based on insight from Psychology that determined that the likelihood of reflecting is tied to one's personality. However, the examples of systems that we reviewed before show that interactive technologies can amplify or enable a user's capacity for reflection. Thus, the related work suggests that there are certain qualities that make interactive technologies more or less effective in supporting reflection. However, these scales were developed with participants of specific professional background in medicine and psychology, consequently being best suited for use in these contexts. Consequently, there is a need for a *'artefact-centric'* instrument that enables the evaluation of technology-supported reflection and measures the qualities of an interactive artefact that lead to reflection. A technology-oriented tool, created in a structured process and validated using various methods (eg. confirmatory factor analysis, test-retest procedure and others) may enable more capable and adequate analysis of reflection in HCI systems, especially within personal informatics field.

3 METHOD

Through exploring the literature we identified concepts related to reflection, which we, in turn, used for generating scale items. A starting point for this inquiry was a study by Li et al. [15] who conceptualised reflection in personal informatics. Their work presents barriers to reflection that should be addressed in the design of interactive systems, including *lack of time* and *effective visualisation of personal data*. Yet, as Baumer et al. [5] noted, there is much to be gained from being explicit in what we mean by reflection. Instead of seeing reflection as a natural consequence or by-product of presenting information to users, he argues that designers of interactive systems should engage with reflection as an important part of a larger set of processes and practices. Earlier work by Baumer [6] further disentangles reflection into three dimensions: breakdown, inquiry and transformation. Breakdown refers to situations of doubt or puzzlement that grab people's attention, and can lead to moments of reflection. Inquiry refers to the process of conscious, intentional inquiry of past experiences in which a person reviews a certain situation. Lastly, transformation exemplifies a change in fundamental assumptions and behaviour. Baumer's stance was in line with earlier work by Fleck and Fitzpatrick [10], who mapped reflection into levels of reflection and conditions for reflection. The levels of reflection consist of five consecutive levels that describe a spectrum of reflective thought, ranging from no reflection to critical reflection. Next to the levels, they also mentioned that creating the right environment for reflection was important, resulting in three conditions for reflection: time, development and encouragement [10].

Next to these dimensions, levels and conditions for reflection, we also used reflection literature from Schön [24] as well as existing scales for reflection [2, 11] as inspiration for the initial set of items.
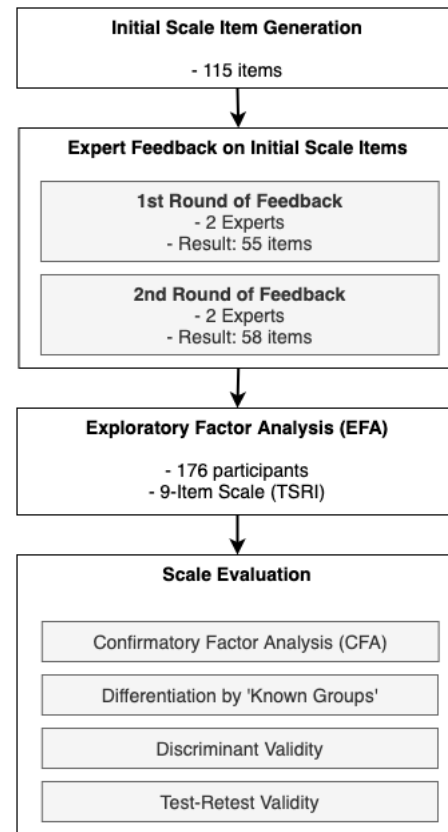

Figure 1: The scale formation process consisted of four phases: (1) the initial generation of scale items, (2) two rounds of expert feedback, (3) Exploratory Factor Analysis, and (4) the evaluation of the scale.

4 SCALE FORMATION

This section describes the process of building the Technology-Supported Reflection Inventory from conceptual considerations to a final list of items.

4.1 Generating Initial Items

Three researchers, each experienced in empirical research in technology for reflection from a variety of academic backgrounds, participated in generating initial items for the scale, which is in line with Boateng et al. [7]. Each researcher first created items, based on related work, on their own. We then arranged a discussion session in which all the generated items were merged and discussed. Through these discussions we removed duplicates. In integrative discussions, we obtained an initial list of 115 items.

4.2 Expert Review

Subsequently, we conducted two rounds of expert reviews to receive feedback on the list of possible scale items. We chose a diverse set of experts to collect broad feedback on the items. The experts in the first round of feedback were a professor in computer science

and a researcher in HCI. They provided feedback by commenting on the list of items and suggesting new items. After collecting the feedback from this first expert review we discussed the list of items, identified items that were problematic and either altered or removed these. This process resulted in a list of 55 items. We then organised a second round of feedback, that followed the same process, yet this time with two researchers in psychology. The feedback of this second expert review was then used to finalise the list of scale items, resulting in a final list of 58 items.

4.3 Survey

After the two expert reviews, we designed an online survey using the Qualtrics XM platform to gather data from participants. We then used this data to perform exploratory factor analysis and item reduction. Since there is a lacking consensus over expected local sample size standards, we aimed for a sample size of 100–200 participants, in line with the development and validation of other scales in HCI [16, 17, 28].

4.3.1 Participants. We recruited a total number of $n = 176$ participants. The participants were recruited using the Amazon Mechanical Turk Service (MTurk) and reimbursed with USD 1. Out of these participants, 93 resided in the USA, 44 in Europe, 38 in Canada and 1 in India. All participants were informed that participation in the study was voluntary and that they could abort the study at any point. We also informed participants that the data would be collected in anonymised form. The survey was conducted online an could be completed in 10 minutes. The average age of participants was $M = 34.1$ ($SD = 10.2$) with 62 of participants identifying as female, 113 as male and 1 participant as non-binary. We recruited a sample size within the expected range, exceeding the ones used for development of related scales in recent studies in HCI [17, 26, 31]. Moreover, our recruitment approach enabled us to reach diverse participant group of various cultural and professional backgrounds, providing more holistic insight than for samples based on student cohorts [2, 11].

4.3.2 Apparatus. We used the Qualtrics platform to design the online survey, the full survey is available in the auxiliary material. To ensure that participants used some form of an interactive technology that can facilitate reflection, we indicated on MTurk that participants needed to use tracking technology (i.e. smartwatch, tracking app) to partake in the survey [1]. The survey consisted of the 55 scale items. Participants scored each item on a 7-point Likert scale, ranging from strongly disagree to strongly agree. Participants were advised to complete the online survey on a computer or tablet, but could also use their smartphone.

4.4 Exploratory Factor Analysis

We conducted factor analysis using a varimax rotation, in which we replicated the approach by Mejia and Yarosh [17]. We then used visual inspection of screen plots to determine the optimal number of factors. After that, we reduced the number of items. We began by removing loadings below 0.40 [7] and items that loaded on multiple factors. We then used an iterative process of removing low loading

---

[1]An overview of the tracking technologies used by our participant is added to the supplementary materials

items and optimising for inter-item reliability in order to further refine the list of items. Current and theoretical Cronbach's alpha coefficients were computed. Our aim was to make the scale as short as possible while maintaining internal consistency and factor structure. This resulted in a loading of three items per factor, and a Cronbach's alpha of $\alpha = 0.78$ for the scale. The theoretical factor model hadcorrect parameters at $TLI = 0.97$ and $RMSEA = 0.04$, cf [7]. Table 1 shows the composition of the scale. We determined the names of the scale dimensions in an iterative discussion with three researchers, mapping them to concepts from our literature review.

5 THE TSRI

The final scale consists of three dimensions that describe the qualities of an interactive technology which supports reflection. Table 1 shows the items included in the theoretical final scale.

5.0.1 Insight. This dimension describes to what extent the interactive technology offers insight to users. It evaluates if the system provides users with information that is insightful to them, and whether it inspires them to look at situations from a new perspective or makes changes to their lives.

5.0.2 Exploration. The second dimension of the TSRI is exploration. It evaluates the ease and enjoyment of exploring personal data in the system. In order for users to reflect on their personal data, it should be easy for users to access and review their data. This dimension can therefore be seen as a necessary condition for reflection on personal data, which echoes findings by Fleck and Fitzpatrick [10].

5.0.3 Comparison. The third dimension of the TSRI measures the social dimension of a refection-supporting interactive system, i.e. the extent to which users can compare themselves to others. It is often argued that current interactive technologies rarely support reflection as a social activity, but instead focus on individual cognitive processes [5, 18]. Yet, reflection can be encouraged by conversations with others [10, 18, 20]. Conversations can provide reflectors with an additional or alternative perspective [10]. This was also shown in a study by Mols [18], 85% of their participants stated that they reflect through conversations with for instance a partner, friend or colleague.

6 SCALE VALIDITY

After building the theoretical composition of the scale, we continued with the evaluation of the TSRI. We conducted Confirmatory Factor Analysis to verify the underlying model. This was followed by a series of tests to check the scale's construct validity and reliability.

6.1 Participants

We recruited $n = 56$ participants via MTurk, using the same approach as in the first survey. The study was conducted online. The reimbursement for completing the survey was USD 0.80. The average age of the participants was $M = 36.6$, $SD = 11.8$, 23 participants identified as female and 33 as male.

Table 1: The TSRI is a nine-item instrument consisting of three dimensions

| Subscale/Item	Factor Loading
Insight, $\alpha = 0.74$	
Q1: Using the system has led to a wake-up call to make changes in my life	0.62
Q2: As a result of using the system, I have changed how I approach things	0.69
Q3: Using the system gives me ideas on how to overcome challenges	0.65
Exploration, $\alpha = 0.72$	
Q4: I enjoy exploring my data with the system	0.68
Q5: The system makes it easy to get an overview of my personal data	0.68
Q6: The system makes it easy to review my long-term personal data	0.65
Comparison, $\alpha = 0.73$	
Q7: I reflect on my data in the system with others	0.65
Q8: The system helps me to discuss my data with others	0.64
Q9: The system makes me think about how my personal data relates with that of others	0.71
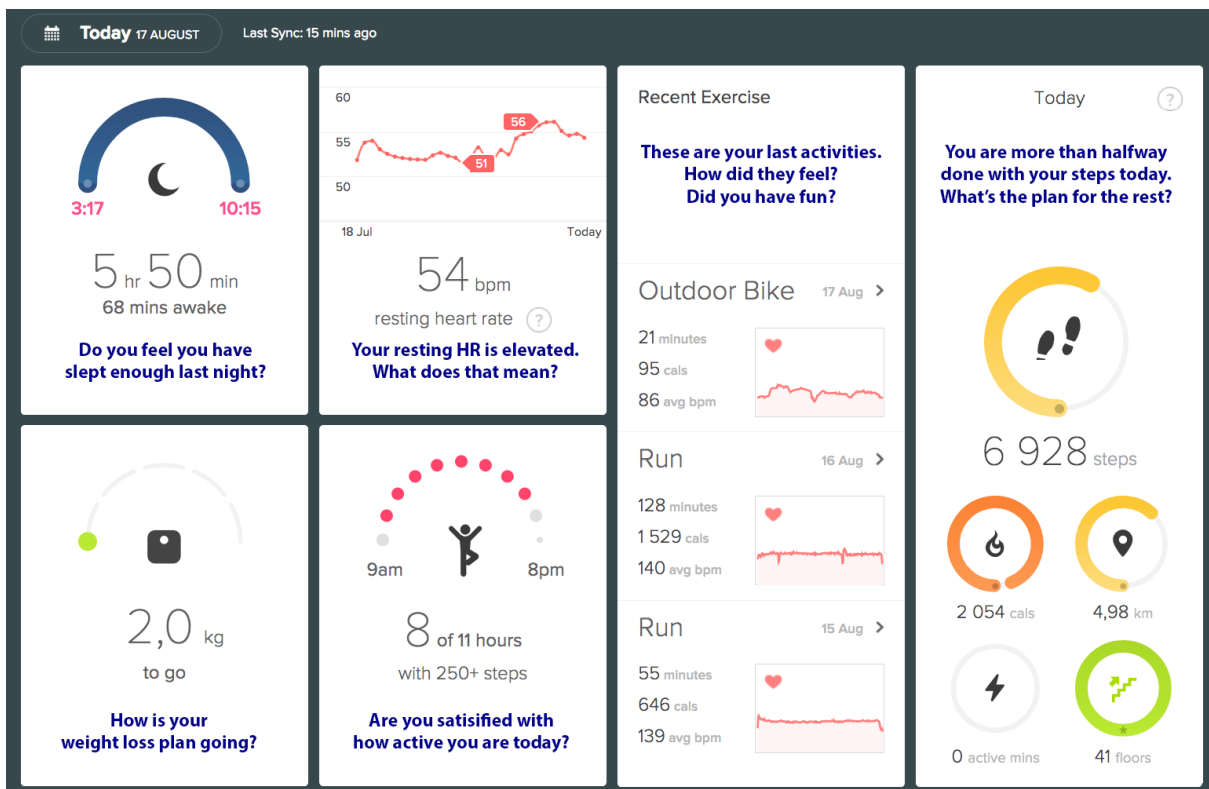
Figure 2: The modified dashboard which we used for testing TSRI's construct validity. The modified dashboard is the standard Fitbit dashboard enriched with questions (which is a known technique for enhancing reflection [18, 20]) regarding the personal data depicted by the dashboard.

6.2 Apparatus

To evaluate the scale, we created a dashboard that showed an overview of personal data from a fitness tracker. We developed two versions; a standard dashboard, and a modified dashboard (see Figure 2). The modified dashboard included questions to enhance reflection. We chose to enrich the dashboard with questions, because this is a known technique for reflection [18, 20]. We assumed the modified dashboard would encourage users more to reflect on their

data when compared to the standard dashboard. We randomly presented each participant with one of the two dashboards and a short description: 'Below, you see some information produced by Brian's fitness tracker. For the rest of the survey, we ask you to become Brian.' Afterwards, we asked participants to 'put themselves in the role of Brian' and to indicate how much they agreed with each item of the (then hypothetical) TSRI about the presented dashboard on a 7-point Likert scale (strongly agree to strongly disagree).

6.3 Confirmatory Factor Analysis

To investigate the correctness of the TSRI factor model, we conducted Confirmatory Factor Analysis (CFA). The resulting Tucker Lewis Index of factoring reliability was $TLI = 0.91$, which is an acceptable fit [7]. We believe that recruiting additional participants and conducting the CFA enabled us to reinforce the reliability of TSRI, even though we recognise this is not a standard approach in developing scales for HCI (eg. [12, 17, 26] did not apply this method).

6.4 Differentiation by 'Known Groups'

Next, we used 'known groups' [7] to contribute to establishing the construct validity of TSRI. To that end, we conducted t-tests to investigate if the TSRI scores differed between the two designs of the tracker dashboard. We found that the modified dashboard scored higher on the full scale as well as on the Insight and Exploration sub-scales. Table 2 shows the results.

6.5 Discriminant Validity

Next, we investigated the discriminant validity of our scale, i.e. determining that the scale was not a derivative of another concept. In our case, we wanted to verify that the TSRI was not a reflection of the users' personal qualities which lead to reflection. To that end, we compared the TSRI with the The Self-Reflection and Insight Scale [11]—a validated instrument for measuring self-regulatory processes related to reflection. We conducted a Pearson Product-Moment correlation test to investigate if the TSRI was correlated with the SRIS. We obtained a result of $r(54) = 0.38$, $p < 0.01$, indicating a weak to moderate correlation. This indicates that the TSRI is conceptually different, but the personality factors measured in SRIS may play a role in its scoring, likely as moderator variables. However, this results also confirms that SRIS can be used effectively to compare interactive technologies, reasonably irrespective of user personalities.

6.6 Test-Retest Reliability

Finally, we checked if TSRI could produce results that were reliable at different points in time, i.e. we tested its temporal stability. We conducted a third online survey for this purpose. In this survey, participants were asked to rate their tracking technology using the TSRI. The full survey and results for the test-retest are available in the auxiliary material.

6.6.1 Participants. We administered the TSRI to a group of $n = 20$ participants, aged $M = 29.3$, $SD = 13.1$, 10 female and 10 male, twice, with a 14-day break in between the studies. Participants were recruited through snowball sampling. The reported sample

does not include participants who did not respond to our invitation to complete the survey for the second time. In line with the first survey, we indicated that participants needed to use tracking technology (e.g. smartwatch, tracking app) to partake in the survey. This ensured that participants used some form of an interactive technology that could facilitate reflection and could form informed opinions.

6.6.2 Results. There is no consensus over what metrics to use to test for test-retest reliability [7]. We used one of Boateng et al.'s suggestions and calculated the intra class correlation coefficient for fixed raters. The result was $\kappa = 0.73$, $p < .01$. The 95% confidence interval ranged from 0.42 to 0.88, According to Koo and Mae al. [14], this indicates moderate reliability with a confidence interval between poor and excellent.

7 DISCUSSION

In this section, we explain how the TSRI can be used in practice and we discuss the limitations of our approach and possibilities for further development.

7.1 Operationalising Reflection

In the tradition of scale development and psychological assessment, content validity is of primary importance. Thus, we followed a structured procedure in our scale development. Multiple psychological assessment instruments were developed, successfully used, and revised using this and similar processes, when theories about the construct (i.e. personal reflection) evolve and new data is acquired (e.g. with the means of our scale) [12]. Our scale, therefore, offers an implicit definition of reflection. This constitutes a possible empirical operationalisation for the HCI field in the face of the lack of an analytically derived definition. Yet, we recognise that there is a need for further understanding and defining technology-supported reflection.

7.2 Using the TSRI

The TSRI can be used in the early stages of designing interactive systems for reflection. Because the scale is relatively short, it can be used for rapid feedback. Using the scale enables the evaluation and comparison of different designs for reflection. However, seeing that a multitude of different definitions of reflection are being used, this might mean that people have contrasting views as to what reflection entails. Thus, it is desirable to compare these different technologies within the same context. By evaluating systems in the same context, the effects of people having different understandings of the term reflection can be minimised. This implies that the scale is not suited to scoring reflection in absolute terms or reflection as an attitude. The TSRI is intended as an artefact-centric instrument that enables researchers and designers to compare different designs relatively from one another.

7.2.1 Scoring and analysis. The TSRI is scored on a seven-point Likert scale from Strongly Disagree (1) to Strongly Agree (7). The items of the scale are balanced over the dimensions and there are no reversed items, which makes the analysis of the scores a straightforward process. The sum of the scores indicates the extent to which an interactive system facilitates reflection. The lowest score on

Table 2: Construct validity assessment differentiation by known groups with t-tests. Bonferroni-corrected p-values are reported.

Scale/Sub-scale	$M_{Modified}$	$SD_{Modified}$	$M_{Standard}$	$SD_{Modified}$	$t(54)$	p
TSRI	46.89	8.86	42.00	8.04	2.16	0.041
TSRI-Insight	17.14	3.04	15.04	2.97	2.62	0.020
TSRI-Exploration	14.86	3.80	12.75	3.69	2.11	0.028
TSRI-Comparison	14.89	4.18	14.21	3.34	0.67	0.477

the scale is 9 and the highest is 63. Higher scores indicate that the artefact offers more support for reflection. The scores offer a transparent and actionable means for designers and researchers to use the TSRI.

7.3 Limitations

There are several limitations to the development of TSRI. In recruiting participants for the study we indicated on MTurk that participants needed to use tracking technology (i.e. smartwatch, tracking app) to partake in the survey. We chose to do that to ensure that participants used some form of an interactive technology that can facilitate reflection. Even though participants used a variety of technologies, with 46% using their smartphones as tracking devices, our considerations were focused on understanding the reflection over self-generated data, collected over time through one's own performance. This might have influenced the formation of the scale, possibly making the scale a better fit particularly for personal informatics systems. We believe that developing related scales, which feature a less data-driven approach is a promising way to understand a wider range of systems which possibly support reflection. Finally, our user samples consisted primarily of Western European and US-based participants. We recognise that the scale, therefore, is constrained within the Western culture and might be prone to cultural biases.

8 CONCLUSION

In this paper, we presented the development and evaluation of the Technology-Supported Reflection Inventory (TSRI). Through a review of the literature, we identified relevant concepts for reflection, which formed the basis for generating scale items. We then built a nine-item scale and TSRI's validity and reliability. We demonstrated the scale's discriminant validity, its ability to differentiate between known groups and test-retest reliability. The TSRI enables researchers and designers to evaluate interactive systems designed for facilitating reflection. The scale can be used to rapidly compare prototypes or assess novel artefacts.

REFERENCES

[1] Riku Arakawa and Hiromu Yakura. 2020. INWARD: A Computer-Supported Tool for Video-Reflection Improves Efficiency and Effectiveness in Executive Coaching. (2020), 1–13. https://doi.org/10.1145/3313831.3376703

[2] Leo C. Aukes, Jelle Geertsma, Janke Cohen-Schotanus, Rein P. Zwierstra, and Joris P.J. Slaets. 2007. The development of a scale to measure personal reflection in medical practice and education. *Medical Teacher* 29, 2-3 (2007), 177–182. https://doi.org/10.1080/01421590701299272

[3] Amid Ayobi, Paul Marshall, and Anna L Cox. 2020. Trackly : A Customisable and Pictorial Self-Tracking App to Support Agency in Multiple Sclerosis Self-Care. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2020), 1–15.

[4] Eric P.S. Baumer, Sherri Jean Katz, Jill E. Freeman, Phil Adams, Amy L. Gonzales, John Pollak, Daniela Retelny, Jeff Niederdeppe, Christine M. Olson, and Geri K. Gay. 2012. Prescriptive persuasion and open-ended social awareness: Expanding the design space of mobile health. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* (2012), 475–484. https://doi.org/10.1145/2145204.2145279

[5] 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. *Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, DIS* (2014), 93–102. https://doi.org/10.1145/2598510.2598598

[6] 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 (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 585–594. https://doi.org/10.1145/2702123.2702234

[7] 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). https://doi.org/10.3389/fpubh.2018.00149

[8] Andrew John Clayphan, Roberto Martinez-Maldonado, and Judy Kay. 2017. A student-facing dashboard for supporting sensemaking about the brainstorm process at a multi-surface space. *ACM International Conference Proceeding Series* (2017), 49–58. https://doi.org/10.1145/3152771.3152777

[9] Daniel A. Epstein, An Ping, James Fogarty, and Sean A. Munson. 2015. A lived informatics model of personal informatics. *UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2015), 731–742. https://doi.org/10.1145/2750858.2804250

[10] Rowanne Fleck and Geraldine Fitzpatrick. 2010. Reflecting on reflection: Framing a design landscape. *ACM International Conference Proceeding Series* (2010), 216–223. https://doi.org/10.1145/1952222.1952269

[11] Anthony M Grant, John Franklin, and Peter Langford. 2002. The Self-Reflection and Insight Scale: A new measure of private self-consciousness. *Social Behavior and Personality* 30, 8 (2002), 821–836. https://doi.org/10.2224/sbp.2002.30.8.821

[12] Stephen N Haynes, David Richard, and Edward S Kubany. 1995. Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological assessment* 7, 3 (1995), 238.

[13] Ellen Isaacs, Artie Konrad, Alan Walendowski, Thomas Lennig, Victoria Hollis, and Steve Whittaker. 2013. Echoes from the past: How technology mediated reflection improves well-being. *Conference on Human Factors in Computing Systems - Proceedings* (2013), 1071–1080. https://doi.org/10.1145/2470654.2466137

[14] 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

[15] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A Stage-Based Model of Personal Informatics Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 557–566. https://doi.org/10.1145/1753326.1753409

[16] Nicola Marsden. 2013. Attitudes towards Online Communication: An Exploratory Factor Analysis. In *Proceedings of the 2013 Annual Conference on Computers and People Research (SIGMIS-CPR '13)*. Association for Computing Machinery, New York, NY, USA, 147–152. https://doi.org/10.1145/2487294.2487326

[17] Kenya Mejia and Svetlana Yarosh. 2017. A Nine-Item Questionnaire for Measuring the Social Disfordance of Mediated Social Touch Technologies. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (12 2017). https://doi.org/10.1145/3134712

[18] Ine Mols, Elise Van Den Hoven, and Berry Eggen. 2016. Informing design for reflection: An overview of current everyday practices. *ACM International Conference Proceeding Series* 23-27-Octo (2016). https://doi.org/10.1145/2971485.2971494

[19] Ine Mols, Elise Van Den Hoven, and Berry Eggen. 2020. Everyday life reflection: Exploring media interaction with balance, cogito & dott. *TEI 2020 - Proceedings of the 14th International Conference on Tangible, Embedded, and Embodied Interaction* (2020), 67–79. https://doi.org/10.1145/3374920.3374928

[20] Jennifer Moon. 1999. *Reflection in learning & professional development.* Routledge-Falmer, New York, NY, USA. 229 pages. https://doi.org/10.4324/9780203822296

[21] Jasmin Niess and Paweł W. Woźniak. 2018. Supporting meaningful personal fitness: The tracker goal Evolution Model. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (2018), 1–12. https://doi.org/10.1145/3173574.3173745

[22] Pablo Ortiz and D. Fox Harrell. 2018. Enabling critical self-reflection through roleplay with chimeria:Grayscale. *CHI PLAY 2018 - Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play* (2018), 365–380. https://doi.org/10.1145/3242671.3242687

[23] Verónica Rivera-Pelayo, Angela Fessl, Lars Müller, and Viktoria Pammer. 2017. Introducing Mood Self-Tracking at Work. *ACM Transactions on Computer-Human Interaction* 24, 1 (2017), 1–28. https://doi.org/10.1145/3014058

[24] Donald A. Schön. 1983. *The Reflective Practitioner.* https://doi.org/10.1542/peds.2005-0209

[25] Petr Slovak, Chris Frauenberger, and Geraldine Fitzpatrick. 2017. Reflective practicum: A framework of sensitising concepts to design for transformative reflection. *Conference on Human Factors in Computing Systems - Proceedings* 2017-May (2017), 2696–2707. https://doi.org/10.1145/3025453.3025516

[26] Hyewon Suh, Nina Shahriaree, Eric B. Hekler, and Julie A. Kientz. 2016. Developing and Validating the User Burden Scale: A Tool for Assessing User Burden in Computing Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16).* Association for Computing Machinery, New York, NY, USA, 3988–3999. https://doi.org/10.1145/2858036.2858448

[27] Anja Thieme, Rob Comber, Julia Miebach, Jack Weeden, Nicole Krämer, Shaun Lawson, and Patrick Olivier. 2012. "we've bin watching you" - Designing for reflection and social persuasion to promote sustainable lifestyles. *Conference on Human Factors in Computing Systems - Proceedings* (2012), 2337–2346. https://doi.org/10.1145/2207676.2208394

[28] Gustavo F Tondello, Rina R Wehbe, Lisa Diamond, Marc Busch, Andrzej Marczewski, and Lennart E Nacke. 2016. The Gamification User Types Hexad Scale. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '16).* Association for Computing Machinery, New York, NY, USA, 229–243. https://doi.org/10.1145/2967934.2968082

[29] Nina Valkanova, Sergi Jorda, Martin Tomitsch, and Andrew Vande Moere. 2013. Reveal-It! The Impact of a Social Visualization Projection on Public Awareness and Discourse. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13).* Association for Computing Machinery, New York, NY, USA, 3461–3470. https://doi.org/10.1145/2470654.2466476

[30] Andrew M Webb, Rhema Linder, Andruid Kerne, Nic Lupfer, Yin Qu, Bryant Poffenberger, and Colton Revia. 2013. Promoting Reflection and Interpretation in Education: Curating Rich Bookmarks as Information Composition. In *Proceedings of the 9th ACM Conference on Creativity & Cognition (C&C '13).* Association for Computing Machinery, New York, NY, USA, 53–62. https://doi.org/10.1145/2466627.2466636

[31] Svetlana Yarosh, Panos Markopoulos, and Gregory D. Abowd. 2014. Towards a Questionnaire for Measuring Affective Benefits and Costs of Communication Technologies. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore, Maryland, USA) *(CSCW '14).* Association for Computing Machinery, New York, NY, USA, 84–96. https://doi.org/10.1145/2531602.2531634