Revising Subjective Evaluation Scales through Collaborative Reflective Re-Scoring: The Case of an Instrument for Assessing User Experience of Mobile Remote Presence

ECCE 2017 Workshop Position Paper

Victor Kaptelinin
Department of Informatics
Umeå University
901 87 Umeå
Sweden
victor.kaptelinin@umu.se

Karin Danielsson
Department of Informatics
Umeå University
901 87 Umeå
Sweden
karin.danielsson@umu.se

Patrik Björnfot
Department of Informatics
Umeå University
901 87 Umeå
Sweden
patrik.bjornfot@umu.se

ABSTRACT

Understanding and assessing various subjective phenomena has long been an important part of HCI research and practice. The most common techniques employed in HCI are interviews, focus groups, and scales, which have their respective strengths and limitations. In the past there have been attempts to combine the strengths of these techniques, for instance by using interviews and focus groups as a basis for generating assessment scales. In this paper we argue that further work is needed to combine assessment scales with more open-ended techniques, and introduce a method, named Collaborative Reflective re-Scoring, which employs collaborative discussion and individual revisions of initial assessment scores, produced by several users, to elicit users' reflections on the reasoning behind their scoring, and thus identify potential problems with the scale-based instrument at hand. The method is illustrated by its use in an empirical study of Mobile Remote Presence.

CCS CONCEPTS

• **Human-centered computing** → User studies • **Computer systems organization** → Robotics

KEYWORDS

Experience evaluation, Mobile Remote Presence, Human-Computer Interaction, Human-Robot Interaction, Collaborative Reflective re-Scoring

1 BACKGROUND

Understanding and assessing various subjective phenomena related to the use of interactive technologies has long been an important part of Human-Computer Interaction (HCI) research and practice. It became even more important with the "experience turn" that the field underwent in the last decade or so [1,4]. This development emphasizes the need for experience assessment instruments, which would make it possible to operationalize and concretize the otherwise abstract notion of experience.

A wide variety of methods and techniques for evaluating experience are currently employed in user research (e.g., [1]), with interviews, focus groups, and scales being the most common ones. Different methods have their respective strengths and limitations. Interviews and focus groups can provide deep insights into users' feelings and thinking in specific contexts, but they are time and effort consuming and the results from using the methods are difficult to compare, summarize, or quantify. Scale-based methods, on the other hand, can be relatively easily applied to obtain data from a large segment of user population, and the data can be efficiently integrated, compared, and analyzed statistically. However, scale-based methods do not reveal the reasons why the informants choose the scores they choose.

Since the respective advantages of, on the one hand, interviews and focus groups, and, on the other hand, assessment scales, are largely complementary, these types of methods are often combined. For instance, in depth interviews can be conducted in order to select a particular set of scales for a survey-type assessment instrument [9].

While interviews and focus groups can indeed provide a valuable input for the development of scales, they do not guarantee the quality of the outcome of such development. The process of constructing scales is not a simple and straightforward one, and insights from interviews and focus groups do not necessarily translate into sufficiently useful, usable, and accurate scale-based instruments.

One way to reveal and address potential problems with a scale-based instrument is to analyze practical applications of the instrument and eventually produce its next version. As the development of some of the most well-known scales, such as NASA-TLX [7] and AttrakDiff [2] suggest, it can be a long process, taking years or even decades to perform. In case of

*

instruments specifically intended for assessing emerging technologies such an evolutionary approach might be too slow. In this paper we propose an alternative approach to the assessment and revision of evaluation scales, named Collaborative Reflective re-Scoring (CRrS), and present an example of how CRrS can be used for an "accelerated development" of a scale-based instrument.

2 CRrS: THE BASIC IDEA

The basic idea of CRrS is revising a scale-based instrument through the following steps: (a) a relatively small group of participants uses it for an initial individual scoring, (b) the group meets and the participants compare their scores with others' scores and explain to each other why certain scores have been assigned, and (c) each member of the group does the scoring again and provides comments on whether, and why, the initial scores have been modified.

The rationale behind CRrS is that collaborative reflection on individual scores creates a natural context for explicating one's intuitive criteria for assigning certain scores, as well as reflecting on how the criteria may be different from those used by other people. By supporting such explication and reflection CRrS helps obtain evidence that can be used to revise current scales and produce an improved version of the instrument in question.

3 APPLYING CRrS FOR ASSESSING USER EXPERIENCE OF ROBOTIC TELEPRESENCE

In our recent study [6] of Mobile Remote Presence (or MRP, which refers to social presence in a local setting via a remotely controlled device combining videoconference capabilities with an ability to move around in the setting [8]), we used CRrS to develop a scale-based instrument for assessing participants' experience of using the technology. While the instrument and its use are presented in other publications reporting the study [3,6], this paper specifically focuses on the approach used to develop the instrument.

The first step in the development of the initial version of the instrument was arranging a session, in which a group of four participants had an opportunity to practically use an MRP system, both as remote and local users. After the session each of the participants was asked to compile a list of adjectives (up to 10) to describe their experience. A combined list of 26 adjectives was generated and converted to a set of 26 semantic differential scales. The instrument was implemented as a survey, a Google form document, which comprised 104 items produced by applying the 26 semantic differential scales to four separate conditions, that is, interacting with the technology as a remote or local user in each of the two versions of the MRP-enhanced environment employed in the study (for details see [6]).

When all participants completed the survey, the group had a 60-minute face-to-face meeting to discuss their scores. The discussion followed the order of items in the survey. The participants, taking turns, explained why they assigned their scores and asked each other for clarification. The discussion was unexpectedly intensive and only 10 items (out of 104) were discussed during the meeting.

After the meetings three of the participants completed the whole survey again. The participants could modify their initial scores, if they wanted to, and were asked to explain their decisions by providing short commentaries. When doing so the participants had access to their own initial scores, as well as the general distribution of scores in the whole group regarding each item of the survey.

The initial set of scales was subsequently revised, mostly as a result of applying the CRrS method. The number of scales in the core set was reduced from 26 to 15, some of the semantic differential descriptions were re-phrased, and the instructions for the participants were further clarified. The revised survey was applied in a larger-scale study, in which it was completed by 32 participants (16 local users and 16 remote users).

4 RESULTS AND CONCLUSION

The empirical evidence obtained in the study includes the themes emerging during the face-to-face discussion, participants' initial and final scores, and the comments provided during re-scoring. The evidence reveals a number of hidden assumptions and diverse interpretations affecting the scores, which need to be taken into account to improve the consistency and accuracy of the instrument. It is worth mentioning that learning about others' scores and scoring criteria made the participants aware of their own, previously implicit, assumptions. Because of space limitations, the evidence cannot be presented here in detail. The authors are going to give an overview of the results at the workshop.

In general, it can be concluded that CRrS offers an effective and efficient way to get an insight into users' interpretation of assessment scales, and thus helps identify potential problems with, and directions for further improvement of, scale-based instruments.

ACKNOWLEDGMENTS

This research was supported by the Swedish Research Council.

REFERENCES

[1] All About UX. n.d.. Available at: allaboutux.org
[2] AttrakDiff. N.d. Available at: http://www.attrakdiff.de/index-en.html
[3] Danielsson, K., and Björnfot. P. (in press). A semantic scale for evaluating the UX of a MRP system. *Proc. of ECCE 2017.*
[4] Hassenzahl, M. 2010. *Experience Design: Technology for All the Right Reasons.* Morgan and Claypool.
[5] Jung, H., Stolterman, E., Ryan, W., Thompson, T., Siegel, M. 2008. Toward a framework for ecologies of artifacts: how are digital artifacts interconnected within a personal life? Proc. NordiCHI 2008. ACM Press: NY
[6] Kaptelinin, V., Björnfot, P., Danielsson, K., and Wiberg, M. 2017. Mobile Remote Presence Enhanced with Contactless Object Manipulation: An Exploratory Study. Proc. of 2017 *CHI Conference Extended.* ACM Press, NY.
[7] Kiselev, A. Loutfi, A. 2012. Using a Mental Workload Index as a Measure of Usability of a User Interface for Social Robotic Telepresence. *Proceedings of the Ro-Man 2012 Workshop in Social Robotic Telepresence.*
[8] Kristoffersson, A., Coradeschi, S., and Loutfi, A. 2013. A review of mobile robotic telepresence. *Adv. in Hum.-Comp. Int.* 2013, Article 3 (January 2013).
[9] Preece, J., Sharp, H., Rogers, Y. 2015. *Interaction Design: Beyond Human-Computer Interaction, 4th Edition.* Wiley.