Multi-Method Evaluation of Adaptive Systems

Christine Bauer
c.bauer@uu.nl
Utrecht University
Utrecht, The Netherlands

ABSTRACT

When evaluating personalized or adaptive systems, we frequently rely on one single evaluation objective and one single method. This remains us with "blind spots". A comprehensive evaluation may require a thoughtful integration of multiple methods. This tutorial (i) demonstrates the wide variety of dimensions to be evaluated, (ii) outlines the methodological approaches to evaluate these dimensions, (iii) pinpoints the blind spots when using only one approach, (iv) demonstrates the benefits of multi-method evaluation, and (v) outlines the basic options how multiple methods can be integrated into one evaluation design. Participants familiarize with the wide spectrum of opportunities how adaptive or personalized systems may be evaluated, and have the opportunity to come up with evaluation designs that comply with the four basic options of multi-method evaluation. The ultimate learning objective is to stimulate the critical reflection of one's own evaluation practices and those of the community at large.

CCS CONCEPTS

• **General and reference** → **Evaluation**; • **Information systems** → **Personalization**; **Recommender systems**; **Evaluation of retrieval results**; • **Human-centered computing** → **HCI design and evaluation methods**.

KEYWORDS

evaluation, multi-methods, personalization, adaptive systems, recommender systems

ACM Reference Format:
Christine Bauer. 2021. Multi-Method Evaluation of Adaptive Systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21), June 21–25, 2021, Utrecht, Netherlands.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3450613.3457122

1 INTRODUCTION

Evaluation is an essential activity in research and development of personalized and adaptive systems. As for most systems, evaluation demands attention in each and every phase through the system's lifecycle—in design and development as well as for continuous improvement while in operation. Accordingly, the evaluation may assess a fully-fledged adaptive system in its entirety, or parts of it—for instance, a model, a technique, or a design facet. The goal of the evaluation task is to observe and measure how well an artifact supports a solution to a defined problem; and, thus, the evaluation activity involves comparing the defined objectives to actual observed results [10].

As personalized and adaptive systems typically integrate various components and can be used for several purposes by a wide set of users with likely different preferences and needs, there is no one-size-fits-all evaluation configuration that could account for all those facets for all forms of adaptive systems or components. Hence, the evaluation configuration has to be adapted to the specific system, scenario, and objective of the assessment. This comes with several (interrelated) challenges: A researcher (team) is challenged to identify and configure an adequate, fitting evaluation design. Further, one single method with one single evaluation metric is frequently not apt to assess the value of an adaptive system in practice. Thus, a thoughtful combination of multiple methods is necessary [2], and a researcher (team) is challenged to identify an adequate combination of method. In addition, it is a challenging task to integrate multiple methods across an entire study so that the results contribute to a comprehensive picture of the evaluated system [4]. This tutorial aims at addressing these challenges.

2 MOTIVATION

The central theme of UMAP 2021 is "Re-Evaluating Evaluation", where the Call for Papers encourages "submissions in all areas that offer a critical analysis of evaluations of personalized systems". This call presupposes (i) knowledge about the historical and current practices of evaluation in the UMAP community, (ii) knowledge and skills in *various* approaches to evaluation, and (iii) the disposition and willingness to critically reflect on potential shortcomings in current evaluation practices (and to come up with viable paths for improvement).

Yet, not every researcher might be knowledgable and skilled in various evaluation approaches. And the lack of adequate knowledge in multiple approaches is regarded as one of the main reasons for the generally low adoption of comprehensive evaluation designs that integrate multiple methods [4]. Likely, this may also an obstacle for a critical "re-evaluation". However, we can leverage the rich community knowledge to improve our evaluation practices. The UMAP community embraces researchers with various backgrounds and skills. On the one hand, we have experienced researchers; on the other hand, researchers early in their academic careers. Some researchers have a long history within the UMAP community and have profound knowledge about evaluation practices in the field; other experienced researchers are new to the domain and may add to the discussion with their experience from other fields. There is a rich basis to learn from each other. This tutorial can give a structured basis for a joint reflection.

UMAP '21, June 21–25, 2021, Utrecht, Netherlands

ACM ISBN 978-1-4503-8366-0/21/06.
https://doi.org/10.1145/3450613.3457122

3 GOALS

This tutorial addresses the call for a critical reflection on evaluation practices in the field. The goals of this tutorial are (i) to demonstrate the wide variety of aspects to be evaluated, (ii) to outline the methodological approaches to evaluate these dimensions (e.g., computational approaches, lab/online experiments with users, field studies), (iii) to pinpoint the blind spots in the evaluation when using only one evaluation approach, and (iv) to demonstrate the benefits of *multi-method evaluation*, and (v) to outline the basic options how to multiple methods can be integrated in one evaluation design.

The most straight-forward learning objectives are that the participants are (a) aware of and familiar with the wide spectrum of opportunities how an adaptive or personalized system may be evaluated, (b) and are able to come up with evaluation designs that comply with the four basic options of multi-methods evaluation. The ultimate learning objective is to stimulate critical reflection of one's on evaluation practices and those of the community at large.

4 A BRIEF OUTLINE OF THE TUTORIAL CONTENTS AND STRUCTURE

The first part of the tutorial provides the conceptual basis, where the participants familiarize with the wide spectrum of opportunities how adaptive or personalized systems may be evaluated. In the second part , the participants have the opportunity to come up with evaluation designs that comply with the four basic options of multi-method evaluation and invites for critical reflection.

Part 1. The tutorial starts with an overview of potentially relevant evaluation goals, considering the perspectives of multiple stakeholders (for details on stakeholder perspectives, see, e.g., [1, 3]), different possible purposes of personalized and adaptive systems (for details on the purpose and value of adaptive recommender systems see, e.g., [8, 9]), and various system properties (for details on the evaluation of properties, see, e.g., [7]). This overview is followed by an outline of the major methodological approaches to evaluate personalized and adaptive systems (i.e., computational or algorithmic approaches, user studies in the lab or in online experiments, and field studies using a real-world system).

Using only a single method frequently leaves us with blind spots in the evaluation; this is illustrated based on a vivid example. After this motivation for multi-method evaluation, the first part of the tutorial proceeds with an introduction to the concept of multi-method evaluation and a discussion of its benefits. Thereafter, the tutorial provides an overview of the four basic options of integrating multiple methods. These evaluation design draw from mixed-methods research and are based on the research designs by Creswell et al. [5, 6] (i.e., the sequential design, the convergent parallel design, the embedded design, and the multi-phase design).

Part 2. The second part of the tutorial is dedicated to practicing and reflection. The participants elaborate multi-method evaluation designs for specific evaluation scenarios for a particular adaptive system. Various possible solutions are discussed. After a discussion of the challenges of multi-method evaluations (for details, see [4]), the tutorial wraps up with a reflection in the plenum addressing, 'Where do we want to go from here?'.

After the tutorial the tutorial slides are made publicly available via *multimethods.info*[1] and *SlideShare*[2].

5 A BRIEF PROFESSIONAL BIOGRAPHY OF THE TUTORIAL PRESENTER

Christine Bauer[3] is an assistant professor at Utrecht University, The Netherlands. Her research and teaching activities are driven by her interdisciplinary background. She holds a Doctoral degree in Social and Economic Sciences (Business Informatics), a Diploma (equivalent to Master) degree in International Business Administration, and a Master degree (MSc) in Business Informatics. In addition, she pursued studies in Jazz Saxophone.

Her research activities center on interactive intelligent systems, where context-adaptivity is a central theme. Recently, she focuses on context-aware recommender systems, and on music recommenders in particular. A core interest in her research activities are fairness in algorithmic decision-making and multi-method evaluations. In her research, she takes a human-centered computing approach, where technology follows humans' and the society's needs.

She is an experienced researcher and holds several best paper awards as well as awards for her reviewing activities. Furthermore, she received the Elise Richter grant by the Austrian Science Fund. Before joining Utrecht University, she was a researcher at Johannes Kepler University Linz, WU Wien, and EC3 (Austria) and University of Cologne (Germany). In 2013 and 2015, she was Visiting Fellow at Carnegie Mellon University (PA, USA). Before starting her academic career, she has built up the field of Licensing New Media at Austria's collecting society AKM, Austria.

She is an experienced teacher in a wide spectrum of topics in computing and information systems—ranging from algorithms to adaptive interactive systems to research methods. Furthermore, she is repeatedly invited as a speaker or panelist at scientific and non-scientific events.

Together with Eva Zangerle, she maintains the website *multimethods.info*[4], where they consolidate resources on multi-method evaluation in research and development of interactive intelligent systems.

ACKNOWLEDGMENTS

I want to express my sincere thanks to Eva Zangerle from Universität Innsbruck who provided valuable insights and expertise that greatly assisted the preparation of this tutorial.

REFERENCES

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *User Modelling and User-Adapted Interaction* (2020). https://doi.org/10.1007/s11257-019-09256-1

[2] Christine Bauer. 2020. Multi-Method Evaluation: Leveraging Multiple Methods to Answer What You Were Looking For. In *2020 Conference on Human Information Interaction and Retrieval* (Vancouver BC, Canada, 14-18 March) *(CHIIR '20)*. ACM, New York, NY, USA, 472–474. https://doi.org/10.1145/3343413.3378015

[3] Christine Bauer and Eva Zangerle. 2019. Leveraging Multi-Method Evaluation for Multi-Stakeholder Settings. In *1st Workshop on the Impact of Recommender*

[1] https://multimethods.info
[2] https://www.slideshare.net/
[3] https://christinebauer.eu
[4] https://multimethods.info

Systems, co-located with 13th ACM Conference on Recommender Systems (ACM RecSys 2019) (Copenhagen, Denmark, 19 September) *(ImpactRS '19, Vol. 2462)*, Oren Sar Shalom, Dietmar Jannach, and Ido Guy (Eds.). http://ceur-ws.org/Vol-2462/short3.pdf

[4] Ilknur Celik, Ilaria Torre, Frosina Koceva, Christine Bauer, Eva Zangerle, and Bart Knijnenburg. 2018. UMAP 2018 Intelligent User-Adapted Interfaces: Design and Multi-Modal Evaluation (IUadaptMe) Workshop Chairs' Welcome & Organization. In *26th Conference on User Modeling, Adaptation and Personalization* (Singapore, 8 July) *(UMAP 2018)*. ACM, New York, NY, USA, 137–139. https://doi.org/10.1145/3213586.3226202

[5] John W. Creswell. 2003. *Research design: qualitative, quantitative, and mixed methods approaches* (2nd ed.). Sage Publications, Thousand Oaks, CA, USA.

[6] John W Creswell and Vicki L. Plano Clark. 2011. *Designing and conducting mixed methods research* (2nd ed ed.). Sage Publications, Los Angeles, CA, USA.

[7] Asela Gunawardana and Guy Shani. 2015. Evaluating Recommender Systems. In *Recommender Systems Handbook* (2nd ed.), Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, Boston, MA, USA, 265–308. https://doi.org/10.1007/978-1-4899-7637-6_8

[8] Dietmar Jannach and Gediminas Adomavicius. 2016. Recommendations with a Purpose. In *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, MA, USA) *(RecSys '16)*. ACM, New York, NY, USA, 7–10. https://doi.org/10.1145/2959100.2959186

[9] Dietmar Jannach and Christine Bauer. 2020. Escaping the McNamara Fallacy: Toward More Impactful Recommender Systems Research. *AI Magazine* 41, 4 (2020), 79–95. https://doi.org/10.1609/aimag.v41i4.5312

[10] Ken Peffers, Tuure Tuunanen andMarcus A. Rothenberger, and Samir Chatterjee. 2007. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems* 24, 3 (2007), 45–77. https://doi.org/10.2753/MIS0742-1222240302