Multi-Method Evaluation: Leveraging Multiple Methods to Answer What You Were Looking For

Christine Bauer
christine.bauer@jku.at
Johannes Kepler University Linz
Institute of Computational Perception
Linz, Austria

ABSTRACT

Research in the field of information retrieval and recommendation mostly focuses on one single evaluation method and one single quality objective. On the one hand, many research endeavors focus on system-centric evaluation from an algorithmic perspective and consider the context of use only to a minor extent. On the other hand, there are research endeavors focusing on user-centric approaches to the design and evaluation of systems. However, algorithmic quality and perceived quality of user experience do not necessarily match. Thus, it is essential for system evaluation to substantially integrate multiple evaluation methods that cover a variety of relevant aspects and perspectives. Only such an integrated combination of methods may lead to a deep understanding of users, their behavior, and experience in their interaction with a system.

This half-day tutorial follows the objective to raise awareness in the CHIIR community concerning the significance of using multiple methods in the evaluation of information retrieval and recommender systems. The tutorial illustrates the "blind spots" when using single methods. It introduces the concept of "multi-method evaluation" and discusses its benefits and challenges. While multi-method evaluations may be designed very flexibly, the tutorial presents broadly-defined basic options of how multiple methods may be integrated in an evaluation design. In group work, participants are encouraged to select and fine-tune a specific design that best matches their research endeavor's purpose.

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, information retrieval, recommender systems, context of use

ACM Reference Format:
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 (CHIIR '20), March 14–18, 2020, Vancouver, BC, Canada.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3343413.3378015

CHIIR '20, March 14–18, 2020, Vancouver, BC, Canada

ACM ISBN 978-1-4503-6892-6/20/03.
https://doi.org/10.1145/3343413.3378015

1 INTRODUCTION

Evaluation for quality is an essential activity in research and system development. Research on software quality exists since the very beginning of software construction [14] and the demand for quality continues to intensify due to our society's increasing dependence on technology [3]. With the advancements of technology, the nature of systems and the context of their use have changed; accordingly, the evaluation objectives have changed too, which also required evaluation efforts and methods to evolve over time [13]. Early research on quality evaluation focused on the internal and development perspective [6]. Now, the users' perspective on quality has become essential because a system not satisfying its users will be less used; which means that it fails in the market [2]. Delivering quality is no longer a competitive advantage but a necessary factor for a system to be successful [2].

The research communities in software engineering and information systems emphasize different methods for improving a system's quality [17, 19]; these perspectives are complementary. Yet, in many research and development endeavors, only one of these two perspectives is adopted, neglecting the respective other one.

In the field of information retrieval and recommendation, for instance, the evaluation procedure in academic research is mainly system-centric and considers the context of use only to a minor extent. Frequently, there is a focus on one single evaluation method and one single quality objective (e.g., prediction accuracy for next-item recommendation). One single method, though, is not able to comprehensively assess all the important aspects for a high-quality system. For instance, system-centric evaluation—which is dominating information retrieval and recommendation research—alone do not comprehensively evaluate a system's quality because—to a large extent—it ignores human aspects in the context of use. For instance, a user's perceived quality [16] depends on the user's context of use in the very moment. User-centric evaluation, in contrast, attempts to put the user in the loop [18] and may involve users interacting with a system or prototype [11] to gather user feedback [5, 10]. Still, neither can user-centric methods alone comprehensively evaluate a system's quality [12]. Similarly, considering that a system typically involves or serves several stakeholders (e.g., providers and consumers) with possibly conflicting interests, it appears unavoidable to employ multiple methods, data sources, or metrics to evaluate for the various aspects [4].

One promising solution for comprehensive evaluation is to integrate and utilize multiple different evaluation methods, covering a variety of relevant aspects and perspectives; such *multi-method evaluation* allows for getting a richer picture of a system's impacts. While this tutorial focuses on the fields of information retrieval and recommendation, multi-method evaluation is a promising strategy for assessing the quality of systems in general.

2 MOTIVATION

For the scope of this work, **the term *multi-method evaluation* refers to an evaluation that integrates at least two different evaluation methods to gather a richer and more integrated picture of a system's quality in the context of use**. The idea to combine different research methods is not essentially new. Mixed methods research [8], for instance, is a research approach where researchers collect and analyze both quantitative and qualitative data within the same study.

Multi-method evaluation is highly related to mixed method research, but broadens the spectrum. While mixed methods research refers to the combination of least one quantitative and at least one qualitative method, the idea of multi-method evaluations is *not* restricted to combining solely (and strictly) quantitative and qualitative methods: multi-method evaluation may integrate several quantitative methods, or several qualitative methods, or combine both.

For multi-method evaluation, we observe a similar phenomenon as for mixed methods research: While attracting considerable interest, it seems that it is rarely brought into practice [1, 15]. From a practical point of view, the reasons for the low adoption of evaluations that leverage multiple methods are manifold, including higher costs (in terms of time and money), a higher level of complexity, a wider set of skills required compared to adopting only one method [7].

This tutorial addresses the skills aspect and aims to contribute the education basis for adopting multi-method evaluation for information retrieval and recommender systems.

3 OBJECTIVES

This half-day tutorial follows the objective to raise awareness in the CHIIR community concerning the significance of using multiple methods in the evaluation of information retrieval and recommender systems. The goals of this tutorial are to point to the "blind spots" in single-method evaluations and their risks involved, introduce to multi-method evaluation, show various approaches how multiple methods may be integrated, provide participants the opportunity to apply these approaches to their individual research endeavors, and receive feedback from their peers.

4 FOCUS

Tutorial contents. This tutorial is planned to be very interactive and tailored to the participants' research endeavors. After raising awareness about the "blind spots" in single-method evaluation and an introduction to multi-method evaluation, participants are encouraged to apply the approaches to their individual research endeavors, discuss their ideas with their peers in small groups, and receive feedback. The structure of the group tasks will strongly orientate on the mixed-methods research designs by Creswell and colleagues [8, 9]. The program of topics and activities is outlined in Table 1 (interactive activities are marked with gray background).

Table 1: Tutorial topics and activities

Topic	Actors
Welcome words and breaking the ice	Facilitator
Motivation, including an illustration of the "blind spots" of single-method evaluations	Facilitator
Introduction to multi-method evaluation, its benefits and challenges	Facilitator
Introduction to "the convergent parallel design"	Facilitator
Task description	Facilitator
Interactive group discussion	Group
Discussion in plenum	Group
Introduction to "the sequential design"	Facilitator
Interactive group discussion	Group
Discussion in plenum	Group & Facilitator
Introduction to "the embedded design"	Facilitator
Interactive group discussion	Group
Discussion in plenum	Group & Facilitator
Introduction to "the multi-phase design"	Facilitator
Interactive group discussion	Group
Discussion in plenum	Group & Facilitator
Summing up and take away message	Facilitator

Learning outcome. The goals of this tutorial are the following:

(i) to raise awareness of the existence and risks of "blind spots" in single-method evaluations,
(ii) to introduce to multi-method evaluation,
(iii) to show various approaches how multiple methods may be integrated,
(iv) to provide participants the opportunity to apply these approaches to their individual research endeavors, and
(v) to receive feedback from their peers.

In the ideal case, a participant leaves the tutorial with a concrete multi-method evaluation design in mind. In the second-best case, a participant knows that he/she/they want to choose another option for a thought-through reason.

5 MATERIALS

After the tutorial, the tutorial slides will be publicly available at *multimethods.info*[1] and on *SlideShare*[2].

It is warmly welcome (but not required) that participants prepare a small task before the tutorial:

[1] https://multimethods.info
[2] https://www.slideshare.net/

Please, reflect on your research goal and the one (or maximum two) research question(s) that you aim to answer.

(1) Write down your research goal in three sentences as concise as possible.
(2) Formulate and write down the research question (or the two questions) that you aim to answer.

Try to be concise, targeted, and "to the point"!

6 TUTORIAL PRESENTER

Christine Bauer[3] is a researcher at the Institute of Computational Perception at the Johannes Kepler University Linz, Austria. In her research, she takes a human-centered computing approach, where technology follows humans' and the society's needs. Her research vision is to leverage intelligent systems and embed them into sociotechnical ecosystems to benefit humans and society. Her main application field are interactive context-adaptive systems. Recently, she focuses on context-aware recommender systems, more specifically context-aware music recommender systems.

Christine's research and teaching activities are driven by her interdisciplinary background. She holds a Doctoral degree in Social and Economic Sciences (Business Informatics) from University of Vienna, Austria, a Diploma (equivalent to Master) degree in International Business Administration from University of Vienna, Austria, and a Master degree (MSc) in Business Informatics from TU Wien, Austria. In addition, she pursued studies in Jazz Saxophone at Konservatorium der Stadt Wien, Austria.

Christine is an experienced researcher. To date, she has authored more than 85 papers in refereed journals and conference proceedings, and holds several best paper awards as well as awards for her reviewing activities. Before joining Johannes Kepler University Linz with her Elise Richter grant by the Austrian Science Fund (FWF), she was a researcher at WU, Austria, University of Cologne, Germany, and the E-Commerce Competence Center, Austria. In 2013 and 2015, she was visiting fellow at the Ubicomp Lab at Carnegie Mellon University, Pittsburgh, PA, USA. Before starting her academic career, she managed and has built up the field of Licensing New Media at Austria's biggest collecting society AKM (Autoren, Komponisten, Musikverleger), Austria.

Christine is an experienced teacher in a wide spectrum of topics in computing and information systems, taught across 10 institutions. Furthermore, she was a speaker at the ACM Summer School on Recommender Systems 2019. At UMAP 2018, she co-organized a full-day workshop on "Intelligent User-Adapted Interfaces: Design and Multi-Modal Evaluation (IUadaptMe)" [7].

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

This research is supported by the Austrian Science Fund (FWF): V579 (https://www.fwf.ac.at).

I further want to thank Eva Zangerle from Universität Innsbruck who provided valuable insights and expertise that greatly assisted the preparation of this tutorial.

[3] https://christinebauer.eu
[4] https://multimethods.info

REFERENCES

[1] Pär J Ågerfalk. 2013. Embracing diversity through mixed methods research. *European Journal of Information Systems* 22, 3 (2013), 251–256. https://doi.org/10.1057/ejis.2013.6
[2] Anas Bassam AL-Badareen, Mohd Hasan Selamat, Jamilah Din, Marzanah A. Jabar, and Sherzod Turaev. 2011. Software quality evaluation: User's view. *International Journal of Applied Mathematics and Informatics* 5, 3 (2011), 200–207.
[3] Jagdish Bansiya and Carl G. Davis. 2002. A hierarchical model for object-oriented design quality assessment. *IEEE Transactions on Software Engineering* 28, 1 (1 2002), 4–17. https://doi.org/10.1109/32.979986
[4] Christine Bauer and Eva Zangerle. 2019. Leveraging Multi-Method Evaluation for Multi-Stakeholder Settings. In *1st Workshop on the Impact of Recommender Systems* (Copenhagen, Denmark) *(ImpactRS '19)*.
[5] Joeran Beel, Marcel Genzmehr, Stefan Langer, Andreas Nürnberger, and Bela Gipp. 2013. A Comparative Analysis of Offline and Online Evaluations and Discussion of Research Paper Recommender System Evaluation. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation (RepSys '13)*. ACM, New York, NY, USA, 7–14. https://doi.org/10.1145/2532508.2532511
[6] Barry Boehm, Sunita Chulani, June Verner, and Bernard Wong. 2007. Fifth Workshop on Software Quality. In *Companion to the Proceedings of the 29th International Conference on Software Engineering (ICSE COMPANION '07)*. IEEE Computer Society, Washington, DC, USA, 131–132. https://doi.org/10.1109/ICSECOMPANION.2007.38
[7] 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). In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore) *(UMAP '18)*. ACM, New York, NY, USA, 137–139. https://doi.org/10.1145/3213586.3226202
[8] John W. Creswell. 2003. *Research design: qualitative, quantitative, and mixed methods approaches* (2nd ed.). Sage Publications, Thousand Oaks, CA, USA.
[9] John W. Creswell and Vicki L. Plano Clark. 2011. *Designing and conducting mixed methods research.* Sage Publications, Los Angeles, CA, USA.
[10] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Transaction on Information Systems* 22, 1 (Jan. 2004), 5–53. https://doi.org/10.1145/963770.963772
[11] Bart P. Knijnenburg and Martijn C. Willemsen. 2015. Evaluating Recommender Systems with User Experiments. In *Recommender Systems Handbook* (2nd ed.), Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, Boston, MA, USA, 309–352. https://doi.org/10.1007/978-1-4899-7637-6_9
[12] Joseph A. Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction* 22, 1 (2012), 101–123. https://doi.org/10.1007/s11257-011-9112-x
[13] Craig M. MacDonald and Michael E. Atwood. 2013. Changing Perspectives on Evaluation in HCI: Past, Present, and Future. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. ACM, New York, NY, USA, 1969–1978. https://doi.org/10.1145/2468356.2468714
[14] José P Miguel, David Mauricio, and Glen Rodríguez. 2014. A review of software quality models for the evaluation of software products. *International Journal of Software Engineering & Applications* 5, 6 (2014), 31–54. https://doi.org/10.5121/ijsea.2014.5603
[15] Guo Chao Alex Peng and Fenio Annansingh. 2013. Experiences in applying mixed-methods approach in information systems research. In *Information Systems Research and Exploring Social Artifacts: Approaches and Methodologies*. IGI Global, Information Science Reference, Hershey, PA, USA, Chapter 14, 266–293. https://doi.org/10.4018/978-1-4666-2491-7.ch014
[16] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-centric Evaluation Framework for Recommender Systems. In *Proceedings of the 5th ACM Conference on Recommender Systems* (Chicago, IL, USA) *(RecSys '11)*. ACM, New York, NY, USA, 157–164. https://doi.org/10.1145/2043932.2043962
[17] Daniel Russo, Paolo Ciancarini, Tommaso Falasconi, and Massimo Tomasi. 2018. A Meta-Model for Information Systems Quality: A Mixed Study of the Financial Sector. *ACM Transactions on Management Information Systems* 9, 3, Article 11 (9 2018), 38 pages. https://doi.org/10.1145/3230713
[18] Markus Schedl, Arthur Flexer, and Julián Urbano. 2013. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems* 41, 3 (2013), 523–539.
[19] Liisa Annikki Von Hellens. 1997. Information systems quality versus software quality a discussion from a managerial, an organisational and an engineering viewpoint. *Information and Software Technology* 39, 12 (1997), 801–808.