IRM PRESS
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.irm-press.com
ITB10761

Chapter 15

Problems and Pitfalls in the Evaluation of Adaptive Systems

Stephan Weibelzahl, Fraunhofer IESE, Germany

Abstract

Empirical studies with adaptive systems offer many advantages and opportunities. Nevertheless, there is still a lack of evaluation studies. This chapter lists several problems and pitfalls that arise when evaluating an adaptive system, and provides guidelines and recommendations for workarounds or even avoidance of these problems. Among other things the following issues are covered: relating evaluation studies to the development cycle; saving resources; specifying control conditions, sample, and criteria; asking users for adaptivity effects; reporting results. An overview of existing evaluation frameworks shows which of these problems have been addressed and in which way.

Evaluation of Adaptive Systems

The demand for empirical evaluations of adaptive systems is getting stronger and stronger. Both researchers and practitioners frequently claim that more empirical studies are required. It seems obvious that empirical research is of high importance for the field, both from a scientific as well as from a practical point of view, because it opens up various advantages and opportunities (Weibelzahl, Lippitsch, & Weber, 2002). For example, empirical evaluations help to estimate the effectiveness, the efficiency, and the usability of a system.

Adaptive systems adapt their behaviour to the user and/or the user's context. The construction of a user model usually requires claiming many assumptions about users' skills, knowledge, needs, or preferences, as well as about their behaviour and interaction with the system. Empirical evaluation offers a unique way of testing these assumptions in the real world or under more controlled conditions. Moreover, empirical evaluations may uncover certain types of errors in the system that would remain otherwise undiscovered. For instance, a system might adapt perfectly to a certain combination of user characteristics, but is nevertheless useless if this specific combination simply does not occur in the target user group. Thus, empirical tests and evaluations have the ability to improve the software development process, as well as the final system, considerably. However, they should be seen as a complement rather than a substitute to existing software engineering methods such as verification, validation, formal correctness, testing, and inspection.

In spite of these reasons in favour of an empirical approach, publications on user modelling systems and adaptive hypermedia rarely contain empirical studies: only about one-quarter of the articles published in *User Modeling and User Adapted Interaction (UMUAI)* report significant evaluations (Chin, 2001). Researchers have been lamenting on this lack frequently (Eklund & Brusilovsky, 1998; Masthoff, 2002), and similar situations have been identified in other scientific areas, too, for instance in software engineering (Kitchenham et al., 2002) or medicine (Yancey, 1996). One important reason for the lack of empirical studies might be the fact that empirical methods are not part of most computer science curricula, and thus, many researchers have no experience with the typical procedures and methods that are required to conduct an experimental study. Moreover, the evaluation of adaptive systems includes some inherent problems and pitfalls that can easily corrupt the quality of the results and make further conclusions impossible.

Given these observations, the objective of this chapter is to provide lessons learned and concrete guidelines to researchers who plan to evaluate their own system. It is supposed to help scientists that have little experience with empirical research to set up studies that fulfil certain quality standards and that do not repeat the errors that have been committed in the studies of the early days of adaptive systems. However, it does not address empirical and experimental issues in general (e.g., proper randomisation, statistical test theory, etc.) and is thus neither a tutorial on statistical methods nor a replacement for in-depth knowledge in empirical methods. It rather illuminates problems that are specific for the evaluation of adaptive systems and offers appropriate recommendations or solutions as far as possible. Having said that, the reader should be aware of the fact that such a list of guidelines must be provisional. Recommendations will inevitably be inappropriate for certain domain-specific problems and approaches. This collection should thus be used as a starting point to consider possible shortcomings of planned study design and data analyses in advance.

The guidelines are supposed to apply for all kinds of user-adaptive systems, that is, all interactive systems which adapt their behaviour to each individual user on the basis of