Problems and Pitfalls in Evaluating Adaptive Systems[1]

Stephan Weibelzahl

National College of Ireland, Mayor Street, Dublin 1, Ireland
sweibelzahl@ncirl.ie

Abstract. Empirical studies with adaptive systems offer many advantages and opportunities. Nevertheless, there is still a lack of evaluation studies. This paper 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 in which way.

1. Introduction

Empirical evaluation of adaptive learning systems is a very important task, as the lack of strong theories, models and laws requires that we do evaluative experiments that check our intuition and imagination. Researchers from various fields have made experiments and published a considerable amount of experimental data. Many of these data sets can be valuable form adaptive learning systems. Still, most of the results are given in a textual form, while structure of these results is not standardized. This limits the practical value of the results. Therefore, if we want to improve the usefulness of the experimental results, it is important to make more formal descriptions of them. The first step toward this goal is creation of the metamodel of empirical evaluation that should identify concepts such as evaluation style, methods and evaluation approaches. This metamodel serves as a conceptual basis form various applications, such as metadescription of experimental data, and creation of experimental data warehouses. Based on this metamodel various tools can work together on creation and processing and comparative analysis of these experimental data.

Given the observation above, 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.

---

[1] This paper is a summary of Weibelzahl, S. (2005). Problems and pitfalls in the evaluation of adaptive systems. In S. Chen & G. Magoulas (Eds.). *Adaptable and Adaptive Hypermedia Systems* (pp. 285-299). Hershey, PA: IRM Press

Adaptive systems adapt their behavior 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 behavior and interaction with the system. Empirical evaluation offers an 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 complement rather than a substitute to existing software engineering methods such as verification, validation, formal correctness, testing, and inspection.

2. Problems and Pitfalls

In spite of these reasons in favor of an empirical approach, publications on user modeling 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. Other problems arise from the nature of empirical work in general. These problems include (Weibelzahl, 2004):

- Formative vs. Summative Evaluation: Often evaluation is seen as the final mandatory stage of a project. While the focus of many project proposals is on new theoretical considerations or some innovative features of an adaptive system, a summative evaluation study is often planned in the end as empirical validation of the results. However, when constructing a new adaptive system, the whole development cycle should be covered by various evaluation studies.
- Allocation of sufficient resources: The fact that evaluations are usually scheduled for the end of a project often results in a radical constriction or even total cancellation of the evaluation phase, because the required resources have been underestimated or are depleted. Empirical work, in particular the data assessment and analysis, require a high amount of personnel, organizational and sometimes even financial resources (Masthoff, 2002). Experiments and real world studies require a considerable amount of time for planning, finding participants, performing the actual data assessment, coding the raw data and statistical analysis.

- Specification of adequate control conditions: Another problem, that is inherent to the evaluation of adaptive systems, occurs when the control conditions of experimental settings are defined. In many studies the adaptive system is compared to a non-adaptive version of the system with the adaptation mechanism switched off (Brusilovsky & Eklund, 1998). However, adaptation is often an essential feature of these systems and switching the adaptivity off might result in an absurd or useless system (Höök, 2000). In some systems, in particular if they are based on machine learning algorithms (e.g., Krogsæter, Oppermann, & Thomas, 1994), it might even be impossible to switch off the adaptivity.
- Sampling strategy: A proper experimental design requires not only to specify control conditions but of course also to select adequate samples. On the one hand the sample should be very heterogeneous in order to maximize the effects of the system's adaptivity: the more the differences between users the higher the chances that the system is able to detect these differences and react accordingly. On the other hand, from a statistical point of view, the sample should be very homogeneous in order to minimize the secondary variance and to emphasize the variance of the treatment. It has been reported frequently that too high variance is a cause of the lack of significance in evaluation studies (Brusilovsky & Pesin, 1998; Masthoff, 2002; Mitrovic & Martin, 2002). For instance, learners in online courses usually differ widely in reading times which might corrupt further comparisons in terms of time savings due to adaptive features.
- Definition of criteria: Evaluating the adaptivity of a system is sometimes seen as a usability-testing problem (Strachan, Anderson, Sneesby, & Evans, 1997). Obviously, usability is an important issue and most adaptive features actually aim at improving the usability. However, there are several aspects of adaptivity that are not covered by usability. For instance, adaptive learning systems usually aim at improving the learning gain in the first place, rather than the usability. The effectiveness and efficiency of other systems are measured in very different ways, as the adaptivity in these systems aims at optimizing other aspects, i.e., the criteria are determined by the system goal and its domain. More details on appropriate evaluation criteria are given below.
- Asking for Adaptivity Effects: In many studies the users estimate the effect of adaptivity (e.g., Beck, Stern, & Woolf, 1997) or rate their satisfaction with the system (e.g., Bares & Lester, 1997; Encarnação & Stoev, 1999; Fischer & Ye, 2001) after a certain amount of interaction. However, from a psychological point of view these assessment methods might be inadequate in some situations. Users might have no anchor of what good or bad interaction means for the given task if they do not have any experience with the 'usual' non-adaptive way. Moreover, they might not even have noticed the adaptivity at all, because adaptive action often flows (or should flow) in the subjective expected way rather than in the static predefined way (i.e., rather than prescribing a certain order of tasks or steps, an adaptive system should do what the user wants to do). Thus, the users might notice and hence be able to report only those events when the system failed to meet their expectations.
- Reporting the Results: Even a perfect experimental design will be worthless if the results are not reported in a proper way. In particular statistical data require special care, as the finding might be not interpretable for other researchers if relevant

information is skipped. This problem obviously occurs in other disciplines and research areas that deal with empirical findings, too. Thus, there are many guidelines and standard procedures for reporting empirical data as suggested or even required by some journals (e.g., Altman, Gore, Gardner, & Pocock, 1983[2]; Lang & Secic, 1997; Begg et al., 1996; Wilkinson & Task Force on Statistical Inference, 1999[3]).

3. Evaluation Approaches

To address at least some of the problems mentioned above, several evaluation frameworks were introduced. These frameworks build upon the idea that the evaluation of adaptive systems should not treat adaptation as a singular, opaque process; rather, adaptation should be "broken down" into its constituents, and each of these constituents should be evaluated separately where necessary and feasible. The seeds of this idea can be traced back to Totterdell and Boyle (1990), who propose that a number of adaptation metrics be related to different components of a logical model of adaptive user interfaces, to provide what amounts to adaptation-oriented design feedback.

The layered evaluation approach (Brusilovsky, Karagiannidis, & Sampson, 2001; Karagiannidis & Sampson, 2000) suggests to separate the *interaction assessment* and the *adaptation decision*. Both layers should be evaluated separately in order to be able to interpret the evaluation results properly. If an adaptation is found to be unsuccessful, the reason is not evident: either the system has chosen the wrong adaptation decision, or the decision was based on wrong assessment results.

Based on these first ideas on layered evaluation, two more frameworks have been introduced that slice the monolithic adaptive system into several layers (respectively stages) that can then be evaluated separately or in combinations (Paramythis, Totter, & Stephanidis, 2001; Weibelzahl, 2001). Recently, these frameworks have been merged, and some validating evidence has been presented (Paramythis & Weibelzahl, submitted). According to this new proposal there are five stages that might be evaluated separately: *collection of input data*, *interpretation of data*, *modeling the current state of the world*, *deciding upon adaptation*, and *applying adaptation*.

In addition, utility-based evaluation of adaptive systems (Herder, 2003) offers a perspective of how to reintegrate the different layers again.

Magoulas et al. (2003) introduced an integration of the layered evaluation approach and heuristic evaluation. Based on existing heuristics that have been used in human-computer interaction (Nielsen, 1994; Chen & Ford, 1998) the authors propose a set of refined heuristics and criteria for every layer. For instance the *acquisition of input data* is evaluated by a heuristic called *error prevention*. It is conducted by checking for typical error prevention techniques (e.g., *data inputs are case-blind whenever possible* or *when learners navigate between multiple windows, their answers are not lost*). In summary, the approach guides the diagnosis of design problems at an early design stage and can thus be seen as a complement to the other frameworks.

---

[2] available at http://bmj.com/advice/

[3] available at http://www.apa.org/journals/amp/amp548594.html

The layered evaluation approach might also be extended by dicing rather than slicing the interaction. Groups of users or even single users might be observed across the layers. Thus, the focus is shifted from the whole sample on one layer to a subset of the sample across layers. For example, the evaluation of an adaptive online course could analyze learners with high and low reading speed separately in order to demonstrate that the inference mechanism works better for one group than for the other. In summary, this perspective might identify sets of (unmodeled but controlled) user characteristics that require a refinement of the user model or at least shape the evaluation results.

It has also been proposed to facilitate evaluation processes through separating design perspectives (Tobar, 2003). The framework integrates abstract levels, modeling issues, traditional concerns, and goal conditions into a so-called extended abstract categorization map which guides the evaluation process. Thus, it addresses in particular the problem of defining adequate evaluation criteria.

This diversity of frameworks and approaches might look a little bit confusing at first glance, but in fact it is a mirror of the current state of the art.

4. Evaluation Criteria

The frameworks and approaches described above provide some guidance concerning adequate criteria for evaluation at each layer. However, the evaluation of the effectiveness and efficiency of a system requires a precise specification of the modeling goals in the first place, as this is a prerequisite for the definition of the criteria. The criteria might be derived from the abstract system goals for instance by using the Goal-Question-Metric method (GQM) (van Solingen & Berghout, 1999), which allows to systematically define metrics for a set of quality dimensions in products, processes and resources. Tobar (2003) presented a framework that supports the selection of criteria by separating design perspectives (see above).

Weibelzahl (2003) also provides an extended list of criteria that have been found in current evaluation studies. For adaptive learning systems obviously the most important and commonly applied criterion is learning gain. However, other general criteria such as learner satisfaction, development of communication or problem solving skills, learner motivation, etc might have to be considered, too. The layered evaluation approach would also suggest evaluating system factors such as the reliability and validity of the input data, the precision of the student model, or the appropriateness of the adaptation decision.

The diversity of these criteria currently inhibits a comparison of different modeling approaches. Future research should aim at establishing a set of commonly accepted criteria and assessment methods that can be used independent of the actual user model and inference mechanism in order to explore the strength and weaknesses of the different modeling approaches across populations, domains, and context factors. While current evaluation studies usually yield a single data point in the problem space, common criteria would allow integrating the results of different studies to a broader picture. Utility-based evaluation (Herder, 2003) offers a way how such a comparison across systems could be achieved.

References

Altman, D., Gore, S., Gardner, M., & Pocock, S. (1983). Statistical guidelines for contributors to medical journals. *British Medical Journal*, *286*, 1489–1493.

Bares, W. H., & Lester, J. C. (1997). Cinematographic user models for automated realtime camera control in dynamic 3D environments. In A. Jameson, C. Paris, & C. Tasso (Eds.), *User modeling: Proceedings of the Sixth International Conference, UM97* (pp. 215–226). Vienna, New York: Springer.

Beck, J., Stern, M., & Woolf, B. P. (1997). Using the student model to control problem difficulty. In A. Jameson, C. Paris, & C. Tasso (Eds.), *User modeling: Proceedings of the Sixth International Conference, UM97* (pp. 277–288). Vienna, New York: Springer.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schultz, K., Simel, D., & Stroup, D. (1996). Improving the quality of reporting randomized trials (the CONSORT statement). *Journal of the American Medical Association*, *276*(8), 637–639.

Billsus, D., & Pazzani, M. J. (1999). A hybrid user model for news story classification. In J. Kay (Ed.), *User modeling: Proceedings of the Seventh International Conference, UM99* (pp. 98–108). Vienna, New York: Springer.

Brusilovsky, P., & Eklund, J. (1998). A study of user-model based link annotation in educational hypermedia. *Journal of Universal Computer Science, special issue on assessment issues for educational software*, *4*(4), 429–448.

Brusilovsky, P., Karagiannidis, C., & Sampson, D. G. (2001). The benefits of layered evaluation of adaptive applications and services. In S. Weibelzahl, D. N. Chin, & G. Weber (Eds.), *Empirical Evaluation of Adaptive Systems. Proceedings of workshop at the Eighth International Conference on User Modeling, UM2001* (pp. 1–8). Sonthofen, Germany.

Brusilovsky, P., & Pesin, L. (1998). Adaptive navigation support in educational hypermedia: An evaluation of the ISIS-tutor. *Journal of Computing and Information Technology*, *6*(1), 27–38.

Chen, S. Y., & Ford, N. (1998). Modelling user navigation behaviours in a hypermedia based learning system: An individual differences approach. *International Journal of Knowledge Organization*, *25*(3), 67–78.

Chin, D. N. (2001). Empirical evaluation of user models and user-adapted systems. *User Modeling and User-Adapted Interaction*, *11*(1-2), 181–194.

Chiu, B. C., Webb, G. I., & Kuzmycz, M. (1997). A comparison of first-order and zerothorder induction for Input-Output Agent Modelling. In A. Jameson, C. Paris, & C. Tasso (Eds.), *User modeling: Proceedings of the Sixth International Conference, UM97* (pp. 347–358). Vienna, New York: Springer.

Eklund, J., & Brusilovsky, P. (1998). The value of adaptivity in hypermedia learning environments: A short review of empirical evidence. In P. Brusilovsky & P. de Bra (Eds.), *Proceedings of Second Adaptive Hypertext and Hypermedia Workshop at the Ninth ACM International Hypertext Conference Hypertext'98, Pittsburgh, PA, June 20, 1998* (pp. 13–19). Eindhoven: Eindhoven University of Technology.

Encarnação, L. M., & Stoev, S. L. (1999). Application-independent intelligent user support system exploiting action-sequence based user modeling. In J. Kay (Ed.), *User modeling: Proceedings of the Seventh International Conference, UM99* (pp. 245–254). Vienna, New York: Springer.

Fischer, G., & Ye, Y. (2001). Personalizing delivered information in a software reuse environment. In M. Bauer, J. Vassileva, & P. Gmytrasiewicz (Eds.), *User modeling: Proceedings of the Eighth International Conference, UM2001* (pp. 178–187). Berlin: Springer.

Herder, E. (2003). Utility-based evaluation of adaptive systems. In S. Weibelzahl & A. Paramythis (Eds.), Proceedings of the Second Workshop on Empirical Evaluation of

Adaptive Systems, held at the 9th International Conference on User Modeling UM2003 (pp. 25–30). Pittsburgh.

Höök, K. (2000). Steps to take before intelligent user interfaces become real. *Interacting With Computers*, *12*(4), 409–426.

Karagiannidis, C., & Sampson, D. G. (2000). Layered evaluation of adaptive applications and services. In P. Brusilovsky & C. S. O. Stock (Eds.), *Proceedings of International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, AH2000, Trento, Italy* (pp. 343–346). Berlin: Springer.

Kitchenham, B., Pfleeger, S. L., Pichard, L. M., Jones, P. W., Hoaglin, D. C., El Emam, K., & Rosenberg, J. (2002). Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering*, *28*(8), 721–733.

Krogsæter, M., Oppermann, R., & Thomas, C. G. (1994). A user interface integrating adaptability and adaptivity. In R. Oppermann (Ed.), *Adaptive user support* (pp. 97–125). Hillsdale: Lawrence Erlbaum.

Lang, T., & Secic, M. (1997). How to report statistics in medicine: Annotated guidelines for authors, editors and reviewers. Philadelphia, PA: American College of Physicians.

Magnini, B., & Strapparava, C. (2001). Improving user modeling with content-based techniques. In M. Bauer, J. Vassileva, & P. Gmytrasiewicz (Eds.), *User modeling: Proceedings of the Eighth International Conference, UM2001* (pp. 74–83). Berlin: Springer.

Magoulas, G. D., Chen, S. Y., & Papanikolaou, K. A. (2003). Integrating layered and heuristic evaluation for adaptive learning environments. In S.Weibelzahl & A. Paramythis (Eds.), *Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems, held at the 9th International Conference on User Modeling UM2003* (p. 5-14). Pittsburgh.

Masthoff, J. (2002). The evaluation of adaptive systems. In N. V. Patel (Ed.), *Adaptive evolutionary information systems.* Hershey, PA: Idea Group Publishing.

Mitrovic, A., & Martin, B. (2002). Evaluating the effects of open student models on learning. In P. de Bra, P. Brusilovsky, & R. Conejo (Eds.), *Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web Based Systems, Málaga, Spain, AH2002* (pp. 296–305). Berlin: Springer.

Nielsen, J. (1994). Heuristic evaluation. Usability inspection methods. New York: Wiley.

Paramythis, A., Totter, A., & Stephanidis, C. (2001). A modular approach to the evaluation of adaptive user interfaces. In S. Weibelzahl, D. N. Chin, & G. Weber (Eds.), *Empirical Evaluation of Adaptive Systems*. *Proceedings of workshop at the Eighth International Conference on User Modeling, UM2001* (pp. 9–24). Freiburg.

Paramythis, A. & Weibelzahl, S. (submitted). A Decomposition Model for the Layered Evaluation of Interactive Adaptive Systems.

Strachan, L., Anderson, J., Sneesby, M., & Evans, M. (1997). Pragmatic user modeling in a commercial software system. In A. Jameson, C. Paris, & C. Tasso (Eds.), *User modeling: Proceedings of the Sixth International Conference, UM97* (pp. 189–200). Vienna, New York: Springer.

Totterdell, P. & Boyle, E. (1990). The Evaluation of Adaptive Systems. In D. Browne, P. Totterdell, & M. Norman (Eds.), *Adaptive User Interfaces* (pp. 161-194). London: Academic Press.

Tobar, C. M. (2003). Yet another evaluation framework. In S. Weibelzahl & A. Paramythis (Eds.), Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems, held at the 9th International Conference on User Modeling UM2003 (pp. 15–24). Pittsburgh.

van Solingen, R., & Berghout, E. (1999). The goal/question/metric method: A practical guide for quality improvement of software development. London: McGraw-Hill.

Weibelzahl, S. (2001). Evaluation of adaptive systems. In M. Bauer, P. J. Gmytrasiewicz, & J. Vassileva (Eds.), *User Modeling: Proceedings of the Eighth International Conference, UM2001* (pp. 292–294). Berlin: Springer.

Weibelzahl, S. (2003). *Evaluation of adaptive systems.* Doctoral dissertation, University of Trier, Trier.

Weibelzahl, S. (2005). Problems and pitfalls in the evaluation of adaptive systems. In S. Chen & G. Magoulas (Eds.). *Adaptable and Adaptive Hypermedia Systems* (pp. 285-299). Hershey, PA: IRM Press

Weibelzahl, S., Lippitsch, S., & Weber, G. (2002). Advantages, opportunities, and limits of empirical evaluations: Evaluating adaptive systems. *Künstliche Intelligenz*, *3/02*, 17–20.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

Yancey, J. (1996). Ten rules for reading clinical research reports. *American Journal of Orthodontics and Dentofacial Orthopedics*, *109*(5), 558–564.