

Advantages, Opportunities, and Limits of Empirical Evaluations: Evaluating Adaptive Systems

Stephan Weibelzahl Gerhard Weber

University of Education Freiburg
Kunzenweg 21
79117 Freiburg, Germany
[weibelza,weber]@ph-freiburg.de

Abstract

While empirical evaluations are a common research method in some areas of Artificial Intelligence (AI), others still neglect this approach. This article outlines both the opportunities and the limits of empirical evaluations for AI techniques exemplified by the evaluation of adaptive systems.

Using the so called layered evaluation approach, we demonstrate that empirical evaluations are able to identify errors in AI systems that would otherwise remain undiscovered. To encourage new evaluations we implemented an online database of studies that are concerned with empirical evaluations of adaptive systems (EASy-D).

1 Advantages: Why Evaluations are needed

Some areas of AI apply empirical methods regularly. E.g., planning and search algorithms are benchmarked in standard domains, and machine learning algorithms are usually tested with real data sets. However, looking at some applied areas such as user modeling, empirical studies are rare. E.g., only a quarter of the articles published in *User Modeling and User Adapted Interaction* (UMUAI) are reporting significant empirical evaluations [4]. Many of them include a simple evaluation study with small sample sizes and often without any statistical methods.

On the other hand, for an estimation of the effectiveness, the efficiency, and the usability of a system that applies AI techniques in real world scenarios, empirical research is absolutely necessary. Especially user modeling techniques which are based on human-computer interaction require empirical evaluations. Otherwise, as we are going to demonstrate in this paper, certain types of errors will remain undiscovered. Undoubtedly, verification, formal correctness, and tests are important methods for software engineering, however, we argue that empirical evaluation—seen as an important

complement—can improve AI techniques considerably. Moreover, the empirical approach is an important way to both, legitimize the efforts spent, and to give evidence to the usefulness of an approach.

2 Opportunities: What we may learn from Empirical Evaluations

According to Cohen [5] empirical methods for AI should answer three basic research questions:

- How will a change in the agent’s structure affect its behavior given a task and an environment?
- How will a change in an agent’s task affect its behavior in a particular environment?
- How will a change in an agent’s environment affect its basic behavior on a particular task?

These questions may be answered by a combination of four kinds of empirical studies: exploratory studies that yield causal hypotheses; assessment studies that establish baselines, ranges, and benchmarks; manipulation experiments to test hypotheses about causal influences; and finally observation experiments (or quasi-experiments) that disclose effects of factors on measured variables without random assignment of treatments [5, 9].

These general and goal defining questions have to be specified in terms of each AI area. As an illustrative example, we outline the opportunities of empirical evaluations for adaptive systems and user modeling. Similar results can be obtained for other AI systems.

The evaluation of adaptive systems can be seen as a layered process where each evaluation layer is prerequisite for the subsequent layers (see KI-Lexikon). Three approaches have been proposed [3, 12, 16] that basically just differ in layer granularity. Thus, we will outline four layers of evaluation of adaptive systems here as introduced by Weibelzahl [16, 17].

Figure 1 shows the four layers: During interaction the adaptive system observes the user and registers certain events or behavior cues (1). Based on these input data abstract user properties are inferred (2). Finally the system decides what and how to adapt (3) and presents the adapted interface to the user (4). Each layer has to be evaluated to guarantee adaptation success.

2.1 Evaluation of Reliability and Validity of Input Data

The first layer evaluates the reliability and the external validity of input data (see KI-Lexikon). Unreliable input data would result in misadaptations. E.g., Spooner and Edwards [13] tried to identify typical dyslexic errors of authors to improve a spell checking system. In an exploratory study they evaluated the stability of certain errors

KI-Lexikon

Layered Evaluation: The Layered Evaluation approach defines several abstract data processing steps within adaptive systems that have to be evaluated in order to guarantee adaptivity success. The evaluation is conducted in layers which means that an successful evaluation of a previous layer is prerequisite for the subsequent layers. E.g., only if the user properties have been inferred correctly it is possible to evaluate different adaptation decisions, because the adaptation decision relies on the user properties.

Objectivity, Reliability, Validity: The quality of observed data may be described in terms of three quality measures. Proper observations are independent of the observer (objectivity), are not biased or distorted by the observation method (reliability), and measure exactly the variable that was intended (validity). As adaptive systems *observe* the user these quality measures are relevant for empirical evaluation.

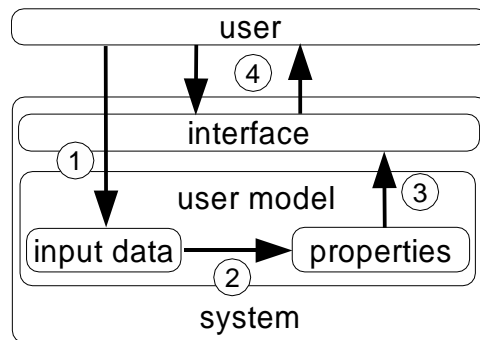


Figure 1: Four layers for the evaluation of adaptive systems.

in time by comparing up to 14 documents of the same author and documents provided by other. Their results suggest that these errors remain stable and might thus serve as input data for further adaptation. If input data turned out to be unreliable, further inferences might be distorted or even impossible.

2.2 Evaluation of Inference

By evaluating the system's inference it is possible to test the inference mechanism in different environments under real world conditions.

Three kinds of studies are used to evaluate the inference. First, exploratory studies can provide empirical grounds for the agent's structure. E.g., Müller, Großmann-Hutter, Jameson, Rummer, and Wittig [11] let their adaptive dialogue system learn the structure of a Bayesian Network from experimental data.

Second, simulations with hypothetical users can prove that certain combinations of input data are processed as expected. E.g., Berthold and Jameson [2] found that their adaptive dialogue system is able to distinct the user groups in the expected way.

Third, in classical experimental settings it is possible to compare the inferences of the system with the real world. E.g., Corbett and Bhatnagar [6] compared the predictions of their adaptive learning system about the students' abilities with the students' actual performance on exercises.

2.3 Evaluation of Adaptation Decision

Even if a system has inferred some user properties there are usually several adaptation possibilities left. Comparing different adaptation decisions (possibly including non-adaptation) estimates the effects of adaptation and may prove the chosen decision to be the most successful. E.g., Weber and Specht [15] compared four different adaptation methods in an adaptive learning system. Each method considered the same user properties but adapted the interface in a different way (i.e., with/without adaptive guiding and with/without link annotation).

2.4 Evaluation of Interaction

The previous layers may show that the system is consistent and infers correct user properties. However, adaptation might still be unsuccessful because users become confused or dissatisfied. Thus, the human-system interaction has to be evaluated as well. Both, objective and subjective measures are relevant. E.g., users might rate the system's usability [7, 14] or the solution quality [1]. Examples of objective criteria for interaction quality include frequency of task success and number of required hints [10].

The examples above emphasize the necessity of empirical evaluations in each of the four layers. It is impossible to detect certain kinds of mis-adaptations that result from biased input data, false inferences, or inadequate adaptation decisions, except for testing the system or parts of the system with real users. Especially usability issues highly depend on empirical research.

3 Limits: Where Empirical Evaluations fail

Empirical research offers many opportunities, however, there are at least two kinds of limitations: on the one hand, errors and pitfalls that are directly related to the layered evaluation approach, and, on the other hand, inherent limitations of empirical research in general.

3.1 General Problems of Empirical Research

Obviously, empirical studies are not a formal proof of a fact. They rather yield, support, or reject hypotheses. However, the results are always afflicted with uncertainty, which can often be expressed in a statistical probability value. Furthermore, for most

statistical tests confidence intervals, test power, and effect sizes are available which should be reported as well.

This hypothesis testing procedure is responsible for an important limitation of empirical research. Empirical studies are very good at identifying design errors and wrong assumptions but they do not suggest new theories or approaches directly. Even an explorative study requires some hypotheses about possible impact factors. Thus, empirical evaluations have to be combined with theoretical grounds to yield useful results.

Not really a limitation but a structural reason why evaluations are currently ignored is the fact that evaluations are not required for publication at international conferences or journals (at least in terms of user modeling). Thus, the empirical part is often scheduled for the end of a project and finally skipped due to lack of time. If publishers and reviewers would demand for empirical evaluations it would soon be an integrated part of research where empirical and theoretical components could stimulate each other. Moreover, AI systems are usually implemented by computer scientists who tend to be less familiar with empirical methods than people with training in human-computer interaction [9].

When evaluating adaptive systems—as opposed to AI systems in general—at least two additional problems emerge: First, defining adequate control groups is difficult for those systems that either cannot switch off the adaptivity, or where a non-adaptive version appears to be absurd because adaptivity is an inherent feature of these systems [8]. Comparing alternative adaptation decisions might relieve this situation in many cases, as this allows to estimate the effect size that can be traced back to the adaptivity itself, but the underlying problem remains: What is a fair comparison condition for adaptive systems.

Second, adequate criteria for adaptivity success are not well defined or commonly accepted: On the one hand, objective standard criteria (e.g., duration, number of interaction steps, knowledge gain) regularly failed to find a difference between adaptive and non-adaptive versions of a system. Usually, these criteria have not been proved to be valid indicators of interaction quality or adaptivity success. On the other hand, subjective criteria that are standard in human-computer interaction research (e.g., usability questionnaires, eye tracking) have been applied to user modeling very rarely. Probably, the effects of adaptivity in most systems are rather subtle and require precise measurement. Recently, a new criterion called behavioral complexity has been proposed [17] that has been designed especially for adaptivity effects but there is still much work to be done on criteria validation.

3.2 Pitfalls and Errors Uncovered by the Layered Evaluation Approach

Having the above in mind there are still several pitfalls that have to be circumvented when conducting evaluations in the different layers.

The evaluation of the reliability of input data relies heavily on a properly selected sample of participants, because retest-reliability and split-half reliability require a sufficient amount of variance in the observed variables. Furthermore, sample selection, sample size, and randomization are important for the subsequent layers as well. Gen-

eralized statements about the inference mechanism are possible only if the observed effect is supposed not to be an artifact of a sample bias [5].

The evaluation of inference will not allow for statements about every possible case including extreme values and special cases as a formal proof would. It will rather test the inference mechanism for external validity and feasibility under real world conditions.

When comparing different adaptation decisions it is possible to select the best one in reference to several criteria. However, there might be unknown or unaccounted adaptation decisions that are even better, because the empirical approach compares of course existing versions only. It might be possible to escape from this limitation by using a human inference mechanism in a so called Wizard of Oz design (or similar approaches) as an additional control condition, because this might account as a benchmark of what adaptation might accomplish in this situation at all [9]. However, this method is applicable only for those kind of systems where humans are actually able to take over the inference processes, as opposed to systems that deal with large amounts of information or complex inferences.

The evaluation of interaction highly depends on a precise and transparent goal setting. Interaction quality can be defined in many different ways, and thus, the result of such an evaluation will never be that "system A is better than system B in general", but only "better in terms of goal X or goal Y".

4 Summary and Future Perspectives

This paper shows that empirical research offers a lot of opportunities that could inspire current research in AI in general and in particular in user modeling. Empirical studies are able to identify errors in AI systems that would otherwise remain undiscovered. However, it has been largely neglected so far.

In order to encourage new empirical evaluations of adaptive systems we implemented EASy-D¹ [18]. This online database contains studies that are concerned with the evaluation of adaptive systems. Each study is categorized in terms of the layer that is evaluated, the criteria that have been used, the function and the adaptation method of the evaluated system(s), statistical methods, and many more dimensions. Researchers who want to evaluate their system get hints about useful criteria that did (or did not) work in previous studies. Proposals of experimental designs and evaluation strategies simplify the planning process.

Moreover, EASy-D could serve as reference for the usefulness of certain inference mechanisms and adaptivity in general. To provide a really useful service to the community the number of registered studies should be expanded considerably. Thus, EASy-D offers an online interface for study submission and everybody is invited to enhance EASy-D with new studies.

¹<http://www.softwareevaluation.de>

References

- [1] Joseph Beck, Mia Stern, and Beverly Park Woolf. Using the student model to control problem difficulty. In Anthony Jameson, Cécile Paris, and Carlo Tasso, editors, *User Modeling: Proceedings of the Sixth International Conference, UM97*, pages 277–288. Springer, Vienna, New York, 1997. Available from <http://www.um.org>.
- [2] André Berthold and Anthony Jameson. Interpreting symptoms of cognitive load in speech input. In Judy Kay, editor, *User Modeling: Proceedings of the Seventh International Conference, UM99*, pages 235–244. Springer, Vienna, New York, 1999. Available from <http://www.um.org>.
- [3] Peter Brusilovsky, Charalampos Karagiannidis, and Demetrios Sampson. The benefits of layered evaluation of adaptive applications and services. In Stephan Weibelzahl, David N. Chin, and Gerhard Weber, editors, *Empirical Evaluation of Adaptive Systems. Proceedings of workshop at the Eighth International Conference on User Modeling, UM2001*, pages 1–8, Freiburg, 2001.
- [4] David N. Chin. Empirical evaluation of user models and user-adapted systems. *User Modeling and User-Adapted Interaction*, 11(1-2):181–194, 2001.
- [5] Paul R. Cohen. *Empirical Methods for Artificial Intelligence*. MIT Press, Cambridge, 1995.
- [6] Albert T. Corbett and Akshat Bhatnagar. Student modeling in the ACT programming tutor: Adjusting a procedural learning model with declarative knowledge. In Anthony Jameson, Cécile Paris, and Carlo Tasso, editors, *User Modeling: Proceedings of the Sixth International Conference, UM97*, pages 243–254. Springer, Vienna, New York, 1997. Available from <http://www.um.org>.
- [7] L. Miguel Encarnação and Stanislav L. Stoev. Application-independent intelligent user support system exploiting action-sequence based user modeling. In Judy Kay, editor, *User Modeling: Proceedings of the Seventh International Conference, UM99*, pages 245–254. Springer, Vienna, New York, 1999. Available from <http://www.um.org>.
- [8] Kristina Höök. Steps to take before intelligent user interfaces become real. *Interacting With Computers*, 12(4):409–426, 2000.
- [9] Anthony Jameson. *Systems That Adapt to Their Users: An Integrative Perspective*. Saarland University, Saarbrücken, 2001.
- [10] Diane Litman and Shimei Pan. Empirically evaluating an adaptable spoken dialogue system. In Judy Kay, editor, *User Modeling: Proceedings of the Seventh International Conference, UM99*, pages 55–64. Springer, Vienna, New York, 1999. Available from <http://www.um.org>.

- [11] Christian Müller, Barbara Großmann-Hutter, Anthony Jameson, Ralf Rummer, and Frank Wittig. Recognizing time pressure and cognitive load on the basis of speech: An experimental study. In Mathias Bauer, Julita Vassileva, and Piotr Gmytrasiewicz, editors, *UM2001, User Modeling: Proceedings of the Eighth International Conference*. Springer, Berlin, 2001. Available from <http://dfki.de/~jameson/abs/MuellerGJ+01.html>.
- [12] Alexandros Paramythis, Alexandra Totter, and Constantine Stephanidis. A modular approach to the evaluation of adaptive user interfaces. In Stephan Weibelzahl, David N. Chin, and Gerhard Weber, editors, *Empirical Evaluation of Adaptive Systems. Proceedings of workshop at the Eighth International Conference on User Modeling, UM2001*, pages 9–24, Freiburg, 2001.
- [13] Roger I. W. Spooner and Alistair D. N. Edwards. User modelling for error recovery: A spelling checker for dyslexic users. In Anthony Jameson, Cécile Paris, and Carlo Tasso, editors, *User Modeling: Proceedings of the Sixth International Conference, UM97*, pages 147–157. Springer, Vienna, New York, 1997. Available from <http://www.um.org>.
- [14] Linda Strachan, John Anderson, Murray Sneesby, and Mark Evans. Pragmatic user modelling in a commercial software system. In Anthony Jameson, Cécile Paris, and Carlo Tasso, editors, *User Modeling: Proceedings of the Sixth International Conference, UM97*, pages 189–200. Springer, Vienna, New York, 1997. Available from <http://www.um.org>.
- [15] Gerhard Weber and Marcus Specht. User modeling and adaptive navigation support in WWW-based tutoring systems. In Anthony Jameson, Cécile Paris, and Carlo Tasso, editors, *User Modeling: Proceedings of the Sixth International Conference, UM97*, pages 289–300. Springer, Vienna, New York, 1997. Available from <http://www.um.org>.
- [16] Stephan Weibelzahl. Evaluation of adaptive systems. In Matthias Bauer, Piotr J. Gmytrasiewicz, and Julita Vassileva, editors, *User Modeling: Proceedings of the Eighth International Conference, UM2001*, pages 292–294, Berlin, 2001. Springer.
- [17] Stephan Weibelzahl and Christoph Ulrich Lauer. Framework for the evaluation of adaptive CBR-systems. In Ivo Vollrath, Sascha Schmitt, and Ulrich Reimer, editors, *Experience Management as Reuse of Knowledge. Proceedings of the 9th German Workshop on Case Based Reasoning, GWCBR2001*, pages 254–263, Baden-Baden, Germany, 2001. Shaker.
- [18] Stephan Weibelzahl and Gerhard Weber. A database of empirical evaluations of adaptive systems. In Ralf Klinkenberg, Stefan Rüping, Andreas Fick, Nicola Henze, Christian Herzog, Ralf Molitor, and Olaf Schröder, editors, *Proceedings of Workshop Lernen – Lehren – Wissen – Adaptivität (LLWA 01); research report in computer science nr. 763*, pages 302–306. University of Dortmund, 2001.

Stephan Weibelzahl received a Diploma in Psychology from the University of Trier and is currently working on his PhD about the "evaluation of adaptive systems" at the University of Education Freiburg (Germany). He is funded by the Graduate School of the Deutsche Forschungsgemeinschaft "Human and Machine Intelligence".

Prof. Dr. Gerhard Weber is a full Professor of Psychology and Dean of Faculty at the University of Education Freiburg (Germany). He developed and implemented the adaptive authoring system NetCoach.