ilmedia

TECHNISCHE
UNIVERSITÄT
ILMENAU

Knauf, Rainer; Jantke, Klaus P.:

Towards an evaluation of (e-)learning systems

Towards an Evaluation of (e-) Learning Systems

Rainer Knauf
Faculty of Computer Science and Automation
Ilmenau Technical University
PO Box 100565, 98684 Ilmenau, Germany
rainer.knauf@tu-ilmenau.de

Klaus P. Jantke
German Research Center for Artificial Intelligence
The Competence Center for e-learning (CCeL)
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
jantke@dfki.de

Abstract

The paper introduces an evaluation approach for learning systems, which is applicable but not limited to e-learning systems. A discussion of current customs in evaluating learning processes reveals some weaknesses of current (not only e-) learning systems, making sophisticated evaluation technologies unsuitable. To overcome these weaknesses, the authors introduced a storyboard concept to represent a learning system's didactic design. This way the subject of evaluation becomes explicit and, thus, assessable to validation technologies. The evaluation approach based on the storyboard concept allows both the communication of general assessments about the system's validity and the indication of the particular weaknesses in the system.

Introduction

The current development of learning processes is characterized by the introduction e-learning systems. This raises many questions of a very general character in the context of learning environments. Besides the request of a proper didactic design, which has been addressed by the authors as well, there is the issue of quality estimation and quality management.

There are only a few publications which address the evaluation issue of e-learning systems. In (Link and Wagner 2004) an analysis of the communication structure reveals basic factors, which determine the quality of computer-mediated communication. Here, the authors consider this issue for different forms and types of uses. Earlier publications (Crystal 2001; Runkehl et al. 1998) tried to adopt conventional research methods for analyzing analogue communication did not make any differentiation of communication forms. However, the results of this research are limited to the quality of communication. This is, indeed, an important factor for the quality assessment of learning environments, but just one. It does not reveal didactic weaknesses like the suitability of a certain material to support the knowledge gain in a certain domain.

What kind of answer do we desire to the issue of a learning system's validity? A first step to this issue is asking for the final purpose of any validity estimation. In fact, this issue has a simple answer: Improving these systems towards better learning results. Some kind of rating of learning results (the learners' skills after some learning activity, e.g.), on the other hand, might be interesting, but does not really help to reveal the particular weaknesses of the learning process. Thus, the process itself should be rated as well.

As a very first step towards an answer to these questions, the authors provide a discussion on several approaches to estimate a learning system's quality. It is intended not to limit these considerations to e-learning systems, but to learning environments in general. Accordingly, the purpose of this discussion is deriving validation technologies for e-learning environments, but not limited to them.

The paper is organized as follows: The next section is an attempt to classify possible evaluation approaches by considering the subject of evaluation and the subjects, who perform the evaluation. Section three considers several learning processes from this perspective. After that, some general requirements to the architecture of learning environments are derived to afford the evaluation of the didactic quality in the fourth section. Section five outlines an approach that meets these requirements and, thus, enables the evaluation of the didactic design. Finally, the entire paper's contribution is summarized.

Evaluation Scenarios for Learning Activities

A first classification of rating approaches can be performed by considering what they evaluate:

1. On the one hand, there is a process-oriented view on this issue by rating the learning activity itself (the quality of a university course by interviewing the students, e.g.).

2. On the other hand, one might have a result-oriented view by rating the learning result, i.e. the skills owned by the learners (by an examination, e.g.).

Both objectives can be followed up by considering the test item (the learning process or the learning result) (a) as a black-box object and (b) as a white-box object.

1. Evaluating the learning process ...

   (a) ... as a black-box object means to provide (maybe criteria-associated) ratings for the entire learning activity, but not for a particular item in this process.

(b) ...in a white-box manner means to consider the sequence of learning scenes and to evaluate, whether or not the particular scene shifts have been the right choice.

2. Evaluating the learning results ...

  (a) ...in a black-box manner means considering only the answer to a question, but not its derivation.

  (b) ...as a white-box object means to consider the solving method and paths used by a student when solving test tasks in an examination.

Another important issue in evaluating learning processes is the source of the assessment. For the really challenging application fields, this source is the (non-explicit and not formally represented) knowledge of human experts. Depending on who provided an evaluation these statements have to be interpreted differently. This is due to the fact, that different humans have different objectives and impressions about the learning success.

For example, there is some indication, that not all students aim at learning as much as they can. Instead, they just aim at receiving good marks. Therefore, when we ask students for an evaluation of a course we run the risk that they don't evaluate it by considering the learning effect, but by the chances to receive good marks instead.

Even students who honestly aim at a maximum learning success, may be blindfolded by the educational environment in general and by personal variables of the educator in particular. (Naftulin et al. 1973), for example, reports about an experiment, in which an actor was trained to teach charismatically but nonsubstantively on a topic about which he knew nothing. In this experiment even experienced educators have been seduced into feeling satisfied with the learning curve and its gained topical knowledge. However, we don't conclude that the learners' satisfaction needs to be excluded when evaluating learning systems. Instead, more sophisticated approaches need to be developed to reveal the real learning success by asking the learners for a rating need to be developed. (Goffman 1959) confirms this need indirectly by admitting that learners fall into the 'illusion trap 'especially when having too little time to evaluate the success of learning. Moreover, the persuasiveness of smart teachers need to be utilized for learning processes (Coats and Swierenga 1972; Rogers 1972). Thus, educators have to be trained to meet the right combination of style and substance.

Since humans are humans, we have to be aware that any rating provided by them is also driven by subjective influences. The only way to obtain a well-balanced rating is to consider the rated objects from different perspectives, i.e. by people with different points of view.

However, two classes (one with subclasses) of evaluators should be distinguished:

3. The evaluation by the teachers, experts, ...

  (a) ... who are involved in the learning process or

  (b) ... who are external experts and

4. the evaluation by the learners (students).

Current Customs

Basic learning activities that aim at constructing very fundamental knowledge and skills take place at (primary, secondary, or high) **schools**. Here, the learning process (see item 1 in the previous section) itself is rarely evaluated. At German schools, for example, there are only two attempts to evaluate the quality of lessons:

- During the study and in the initial phase of a (future) teacher's work, some mentor sits in the back of the class and rates the teacher's classes.

- If somebody complains about the classes (parents or pupils, e.g.) or the examination results are too bad, the corresponding teacher might be contacted by his/her officer and might be asked to change the manner of teaching.

As long as nobody complains and the examination results are not too far from normality, the quality of the learning process is never evaluated at German schools.

The learning result (see item 2), on the other hand, is evaluated all the time: The pupils have to provide answers to the teacher's questions, to pass tests during the classes, to pass examinations after the classes, and so on.

These learning results are checked exclusively by teachers, i.e. 'topical experts'(see item 3) and (maybe besides a few exceptions) by their own teachers (see item 3a). An evaluation by the pupils (see item 4) usually doesn't take place.

This evaluation setting shifts, when we consider higher level learning activities as those in German **universities**, for example (Dawideit et al. 2003). Currently, there are some first attempts to evaluate the learning process (see item 1) permanently:

- Formally, to become a professor, one has to pass the so-called habilitation process. Besides the high level research skills the candidate is checked for teaching skills as well. This is a very old custom in Germany: For more than 100 years there is an examination called *facultas docendi*, which is a part (unfortunately, a very small one) of the habilitation process. This evaluation considers test classes and an official test tutorial, i.e. the teaching process as a white-box (see item 1b).[1] This rating is performed by an external expert panel (see item 3b).

- Additionally, it becomes more and more a custom, that students rate their classes (see item **??**). They do not rate particular didactic decisions, but the entire class by considering various criteria. This is a black-box rating (see item 2a). Unfortunately, this rating is quite disputed and does not really have an impact to the university teacher.

The learning results (see item 2) are also evaluated all the time. Students have to submit home works, to give talks about their work, to pass examinations during and after the course and so on. These results are usually reviewed by considering the particular problem solving steps, i.e. in a white-box manner (see item 2b) by their own teachers (see item

[1]However, in practice there are many exceptions to this rule: University classes are also taught by people, who did not take part in this examination process (which doesn't necessarily mean, that they are worse), and at so-called *Universities of Applied Sciences* such a qualification is not requested at all.

3a). The higher the academic level, the more external experts are employed to provide the reviews (see item 3b). A Ph.D. panel, for example, usually has to consist of a minimum number of professors from other universities.

In fact, there are many other institutions (than schools and universities) which perform teaching activities. To pick up an example of learning a skill which might become a threat if not present, let's consider **driving schools**. Here, the result of learning is evaluated (see item 2), not the process. At a first view, the theoretical skills are examined by the own teachers. Since both the questionnaire and the evaluation process are a Federal standard in Germany, this examination has to be classified as an evaluation by external experts (see item 3b). The practical examination is performed by a panel which also includes external experts (see item 3b).

General Requirements

Because of the very different customs in different institutions, different countries and culture areas, different subjects to be learnt and so on it is difficult to derive some **general rules** on

- what is the subject of evaluation (the process or the result),

- whether this subject is considered in a black-box – or white-box manner, and

- who performs the evaluation (involved or external persons, experts or students).

We are aware, that the following statements are quite vague, but they might be helpful, when deriving evaluation scenarios for e-learning environments:

(A) Unfortunately, the process of learning is rarely evaluated. Since the 'dramaturgy of the learning process ' is a significant variable in the evaluation of teaching effectiveness (Gage 1963), its explicit representation provides a basis for its evaluation. To point out the particular weaknesses of this process by a validation technology, this process needs to be the subject of evaluation.

(B) The evaluation is rarely performed by the learners themselves. At least adult and mature learners might also provide useful hints to improve the quality of learning processes. To avoid the trap to feel satisfied with the learning result without having learned anything (Naftulin et al. 1973), the learners' ratings need to be acquired by technologies beyond reflecting their illusion of having learned. Since learning is an interactive process of constructing topical knowledge, all parties of this interaction are qualified to assess the quality of this process. Ignoring the learners' didactic experience wastes a valuable source of validation knowledge.

(C) The higher the academic level of a skill to be learnt or the higher this skill is mission critical, . . .

(C1) . . . , the more the evaluation should shift from a black-box towards a white-box evaluation.

(C2) . . . , the more external experts need to be included to provide ratings.

To meet the requirement (**A**), ratings of the learning process have to be attached to particular representations of both (a) the topical knowledge to be learnt and (b) the didactic knowledge used to support the learning process. For this purpose, an explicit representation of both is essential:

(a) Learning environments so far are often characterized by dividing the knowledge to be learnt into a hierarchy of learning units, which can be considered as material, which is helpful to construct topical knowledge in the learner's brain.

The hierarchy itself, i.e. the structuring of these units, is formal. The content of the atomic (not composed) elements can by anything and is not necessarily formal. In fact, for most interesting domains it is informal.

(b) The didactic knowledge, on the other hand, is usually not represented explicitly. The didactic experience might be available as 'hidden knowledge' of human teachers and might have been evaluated during the initial qualification of these teachers, but in this form of appearance it can't be evaluated permanently so far.

To bridge this gap, the authors introduced a concept for the didactic design of (not only e-) learning processes (Jantke and Knauf 2005), which is outlined in the following section.

Didactic Design through Storyboarding

The authors' storyboard concept (Jantke and Knauf 2005) is built upon standard concepts which enjoy an appealing visual appearance: graphs. Here is the core notion:

A storyboard is a graph with annotations to its nodes and edges. Nodes are scenes or episodes; the edges specify transitions between them. Scenes are elementary and may be implemented in different ways. Episodes are composite and are described by subgraphs. Key annotations to nodes specify actors and locations.

The concept may be refined by numerous additions as listed below. Note that all the following supplements are not really necessary. Many of them are implicit in the general concept. We discuss those details only for the readers' convenience, to become a little more familiar with our ideas, aims and intuition. In *italics* we provide details about our current technological representation of storyboards in Visio^{TM} (Martin 2002; Martin 2003). Readers may use any other appropriate tool at hand.

- Because those nodes that are called episodes may be expanded by subgraphs, storyboards are hierarchically structured graphs by their very nature. *Double clicking on an episode opens the corresponding subgraph on a separate sheet.*

- Comments to nodes and edges are intended to carry information about didactics. Goals are expressed and variants are sketched. *Clicking to a comment opens a window with the text, including author information and date.*

- As far as it applies to a node, educational meta data (SCORM, LOM, . . .) may be added. *Visio built-in object properties are used to represent general information and meta data.*

- Edges are colored to carry information about activation constraints and any variants of their adaptive availability. Certain colors may have some fixed meaning like usage for certain educational difficulties. *Clicking on edges opens didactic comments and meta data for adaptive behavior.*

- Actors and locations inclusive those in the real world are assigned to elementary nodes only. *Through programming, actor and location information may be propagated automatically.*

- Certain scenes represent documents of different media types like pictures, videos, PDFs, Power Point slides, Excel Tables, . . . *Double clicking on a scene opens the media object in a viewer, i.e. plays the film, e.g.*

Figure 1: The Top Level Storyboard of a Data Mining Course

For illustration, figure 1 shows a top level storyboard that has recently been designed by the authors for a course on Data Mining. Figure 2 shows an atomic scene from the subgraph behind the episode *Competitive Exercises* of this storyboard with a comment to an edge (on the left) and a node (on the right). In a more detailed storyboard, this scene may be turned into an episode and, thus, be further refined subsequently. Therefore, storyboarding for e-learning is a
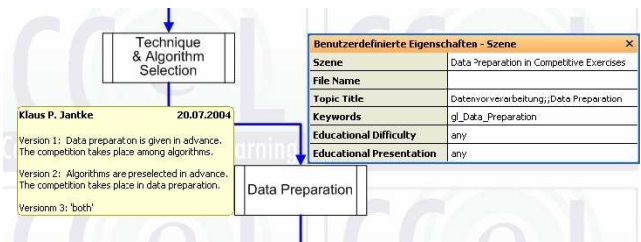

Figure 2: An Atomic Scene with annotations from the Storyboard

processes that might never end. One may always take a scene, declare it an episode and continue in-depth design.

Clearly, the sophistication of storyboards can go very far. The concept allows for deeply nested structures involving different forms of learning, getting many actors involved and permitting a large variety of alternatives. Though this is possible, in principle, the emphasis of this concept – driven by the goal of dissemination – is on simple storyboards designed quickly.

In our storyboarding practice, discussions begin frequently with a top level storyboard of only about half a dozen nodes. Discussing their arrangement is a first step toward didactic design.
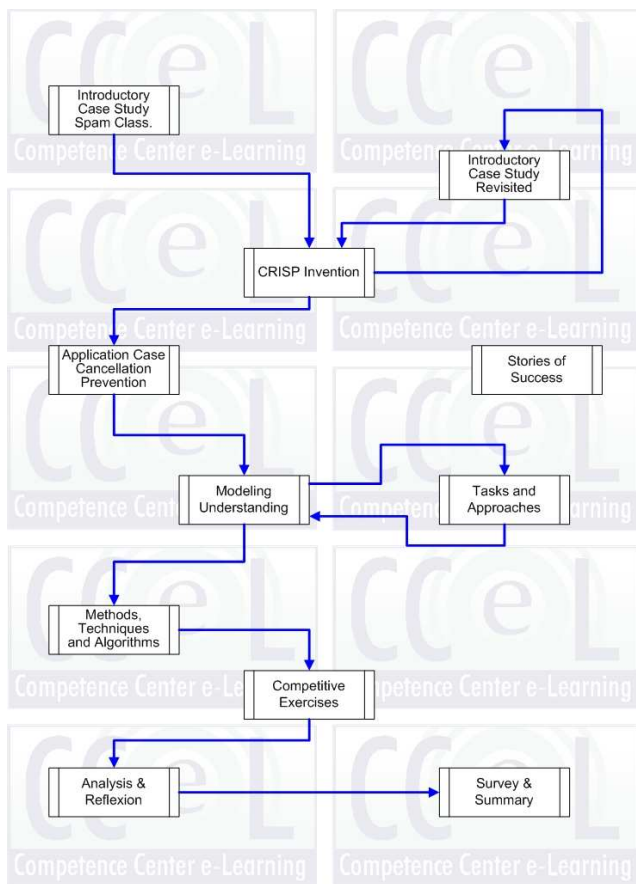
An Evaluation Approach

This concept to organize both the material that helps to construct topical knowledge in the learner's brain and the didactic knowledge provides a firm basis to evaluate learning processes. It allows the attachment of ratings to particular representations of both and, thus, to point out the particular weaknesses of a learning process.

A basic pre-condition is the availability of humans who provide ratings. Again, no potential source of ratings should be wasted: Whoever came in contact with a learning process under evaluation (learners, teachers, external experts, observers, . . .) is a valuable source of validation assessments.

Having a concept like the storyboard as sketched above, any learning process is characterized by a particular path within each of the nested graphs. This is the key to evaluate the learning process. Whenever some can rate the success of process, this rating is attached to the nodes and edges of the traversed path.

- A rating of a node, which represents a scene, reflects the quality of its topical content respectively the need to refine the didactics by constructing a subgraph to it.

- A rating of a node, which represents an episode, reflects both the quality of its content and its didactic design represented by its subgraphs.

- A rating of an edge within this graph reflects the validity of the didactic decision to shift from source node to the destination node.

Since the ratings of the concept introduced here need to be computable, we did not adopt any rating system of a school, a university, or any other learning institution. Instead, we propose to rate both the nodes and the edges by a rational number within the range 0 through 1 (both included),

whereas 1 is the very best rating and 0 indicates that the node or edge is completely insufficient.

Before the initial use of a learning system represented with the storyboard concept, there is only one source of (human) ratings: The author(s) of the storyboard. Since they usually don't intend to make bad job, we need to rate every element of the initial storyboard by 1.

For simplification[2], every person (learner, teacher, external expert, ...) who is asked to provide an evaluation, can express it by evaluating it as being 'good' or 'bad'.

Again, the evaluation process needs to be attached to a *particular* learning process, i.e. to a path within each of the nested graphs in the storyboard, which have been used in this particular process. Thus, any learning process, i.e. the traversed path, must be recorded in a log file. The success or failure of a learning process is the success or failure of the traversed path in the storyboard, independently from the manner of deriving the evaluation (by a written and/or oral test evaluated by teachers or external experts, by a learner's rating of his/her success, by the parents' rating of their kids' success in school, ...).

Whenever an evaluation to such a process (and, thus, a path in each of the nested graphs) is provided, the current ratings of the nodes and edges of these particular paths needs to be adjusted

- towards a better rating in case the entire process has been considered as 'good' respectively
- towards a worse rating in case the entire process has been considered as 'bad'.

The authors suggest the principle of *exponential smoothing* to adjust a rating upward or downward.

The influence of a new evaluation statement (like 'good' or 'bad' in our setting of the evaluation concept) needs to be quantified in advance and according to the topical domain, the learning purpose, the learning conditions, the expected number of available evaluations, and so on.

Moreover, this influence can be requested to be different according to the source of the evaluation statement: the teacher, the learner, an external expert, an individual, who was not involved in the process, and so on.

In the suggested approach (exponential smoothing) this influence is quantified by a (usually quite small) rational number w ($0 < w << 1$). The rule to adjust the ratings after the availability of a new (let's say the i-th) evaluation is

$$r_i := (1 - w) * r_{i-1} + w * eval$$

with

- r_i, r_{i-1} being the rating after the i-th respectively $(i-1)$-th evaluation,
- $eval$ being the new (i-th) evaluation: $eval := 1$, if this the evaluation 'good' and $eval := 0$ otherwise, and
- w being the influence of the new evaluation according to its source.

───────────

[2]The rating system may be refined after gaining some experience with its application.

To clarify the role of the influence factor w, it might be helpful to note, that there is the following correlation between w and the number n_h of evaluations $eval := 0$, that is necessary to halve a node's or edge's rating:

$$
\begin{aligned}
(1 - w)^{n_h} &= 0.5 \\
n_h &= \frac{log\ 0.5}{log(1 - w)} \\
w &= 1 - 10^{\frac{log\ 0.5}{n_h}}
\end{aligned}
$$

After applying this approach over a sufficient number of learning sessions the ratings for the nodes and edges reflect the success of their use and, thus, the validity of the performed learning processes. This rating approach outdates historic evaluations stepwise exponentially and promotes the most recent evaluations.

The degree to which expand the success of performed learning processes really reflects the learning system's validity, depends (among other things) on the coverage of these processes with respect to the entire network of possible paths that are represented in the system. To communicate an estimation of the entire learning system's validity, the following parameters could be helpful:

- The average rating of the scenes reflects the quality of its topical content respectively the need to refine the didactics by constructing a deeper nesting of the graphs.
- The average rating of the episodes reflects both the quality of their content and their didactic design.
- The average rating of the edges reflects the quality of the didactic decisions to shift episodes and scenes.

One might argue, that the didactic design of a learning process has some intended features, which might lead to 'bad marks' when applying this approach:

- Nodes and edges, which are visited in case of not having learnt a content successfully and point to some repetition process. These nodes might rarely been visited.
- Nodes and edges, which are intended as alternative ways to reach the same learning objective share the same visiting frequency as a node with no alternative.
- Nodes and edges, which are default-shifts for unexpected events, are usually never visited.

Generally, these concerns are unnecessary. The fact, that a node or edge is visited rarely, does not necessarily mean that it receives 'bad marks'. The essential issue is, which rating a node or edge receives, but not how often it receives a rating. Since the initial rating for each element of the storyboard is 1, a never visited element keeps this rating all the time. This is natural, because there is only one (implicit, but positive) rating available: The storyboard developer's assumption that he/she made a good job, since he/she won't produce bad storyboards by intention.

A characteristic feature of this approach can be used to support the system's refinement:

- The worst rated scenes point out topical weaknesses and/or the need to refine the didactics by constructing subgraphs. To distinguish topical from didactic weaknesses, we propose to ask topical experts for the first. By receiving a confirmation of topical correctness, we can conclude a didactic weakness.

- The worst rated episodes can point out both a bad content behind and/or a bad didactic design. A more detailed description of the particular weaknesses can be derived by jumping into the subgraphs and considering the ratings in it.

- The worst rated edges point out bad didactic decisions to shift from source node to the destination node.

Summary and Conclusion

The paper introduces an evaluation approach for learning systems, which is applicable but not limited to e-learning systems. A discussion of current customs in evaluating learning processes leads to requirements, which make learning systems assessable for an evaluation of their topical and didactic quality. To meet these requirements, the authors introduced the storyboard concept for didactic design, which turned out to be a firm basis for the learning system's evaluation. Storyboarding is the key to make the particular subjects of evaluation explicit. The evaluation approach is based on this concept and allows to communicate general assessments about the system's validity as well to point out its particular weaknesses.

Further developments of this approach are directed towards

1. refining the evaluation scale,

2. systematically constructing learning objectives, which enforce desired paths in the storyboard and, thus, ensure a sufficient coverage of storyboards by using them as test processes, and

3. deriving more sophisticated validity statements from the ratings, which make them really useful for the learning system's refinement.

The intention behind the latter item is to identify design patterns and derive a rating measure from the ratings of their components. A typical design pattern is shown in figure 3. The identification of typical patterns can be performed by
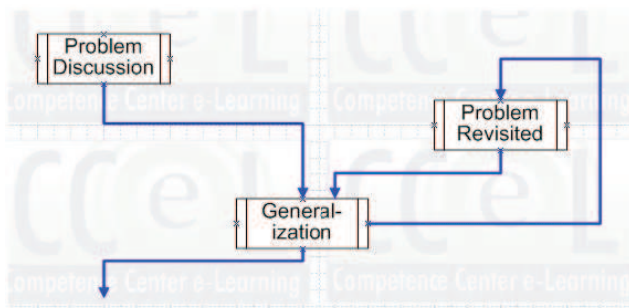

Figure 3: A Typical design Pattern

analyzing particular successful storyboards. This is, in fact, the first step towards exploring and learning (new) general didactic knowledge. Indeed, it is a conceptual challenge, but there is no need to develop a particular tool, since Visio^{TM} (Martin 2002; Martin 2003) provides the opportunity to include any analysis technique and any technique to compute a rating for such patterns.

References

Coats, W.D.; Swierenga, L. 1972. Student Perceptions of Teachers. A Factor Analytic Study. *Journal of Educational Research*, 65:357–366, April, 1972.

Crystal, D. 2001. *Language and the Internet*. Cambridge: University press, 2001.

Dawideit, T.; Merkt, M.; Petersen, K.; Schädlich, B.; Schulmeister, R.; Windisch, A. 2003. '*Master of Higher Education 'Modellversuch zur didaktischen Professionalisierung von Hochschullehrenden*. UniversitätsVerlag Webler, 2003.

Gage, N.L. (Ed) 1963. *Handbook of Research on Teaching*. New York: Rand McNally, pp. 506, 464, and 481, 1963.

Goffman, E. 1959. *The Presentation of Self in Everyday Life*. New York: Doubleday, 1959.

Jantke, K.P.; Lange, S.; Grieser, G.; Grigoriev, P.; Thalheim, B.; Tschiedel, B. 2004. *Work-Integrated E-Learning – The DaMiT Approach*. Accepted paper. Proc. of the 29th Internat. Scientific Colloquium, September 27–30, Technical University of Ilmenau, Germany, 2004.

Jantke, K.P.; Knauf, R. 2005. Didactic Design through Storyboarding: Standard Concepts for Standard Tools. Accepted paper. Proc. of the Winter Internationals Symposiums of Information and Communication Technologies, First International Workshop on Dissemanitation of E-Learning Technologies and Applications, January 3-6, 2005, Cape Town, South Africa, 2005.

Link, L.; Wagner, D. 2004. Methods for Evaluating CMC in Online Learning in Higher Education. Auer/Auer (eds.): *International Conference on Interactive Computer Aided Learning (ICL 2004)* (CD-ROM), Sept. 29 – Oct. 1, 2004, Villach, Austria, ISBN 3-89958089-3, 2004.

Martin, R. 2002. *Visio 2002 für Anwender*. Software & Support Verlag, 2002.

Martin, R. 2003. *Visio programmieren*. Software & Support Verlag, 2003.

Naftulin, D.H.; Ware Jr., J.E.; Donnelly, F.A. 1973. The Doctor Fox Lecture: A Paradigm of Educational Seduction. *Journal of Medical Education*, vol. 48, July 1973, pp. 630–635, 1973.

Rogers, C.R. 1972. Bringing Together Idea and Feelings in Learning. *Learning Today*, 5:32–43, Spring, 1972.

Runkehl, J.; Schlobinski, P.; Siever, T. 1998. Sprache und Kommunikation im Internet: Überblick und Analysen. Opladen, Wiesbaden: Westdeutscher Verlag, 1998.

Tergan, S.-O.; Schenkel, P. 2004. *Was macht e-Learning erfolgreich? Grundlagen und Instrumente der Qualitätsbeurteilung*. Springer, 2004.