A Classification of Evaluation Methods for Intelligent Tutoring Systems

Arif Iqbal[*], Reinhard Oppermann[**], Ashok Patel[***] and Kinshuk[**]

[*]University College London UK, [**]GMD-FIT Germany, [***]De Montfort University UK

Summary

Evaluation of intelligent tutoring systems (ITS) is an important area of research in current educational practices. There are many evaluation methods available but the literature does not suggest any clear guidelines for an evaluator – normally an educator – which methods to use in particular contexts. This paper proposes a classification of evaluation methods to simplify the selection task. The classification is based on two primary questions relating to the target of evaluation and learning environment in which the evaluation would be pursued. The classification is hoped to help in improving quality of computer based education by providing a practical and to the point way of selecting the appropriate evaluation methods for intelligent tutoring systems.

Zusammenfassung

Die Evaluation intelligenter, tutorieller Systeme (ITS) ist ein wichtiges Forschungsfeld im Bereich computer-gestützten Lehrens und Lernens. Bei einer Vielzahl vorhandener Methoden finden sich in der Literatur aber keine klaren Richtlinien für einen Evaluator – üblicherweise ein Lehrender – dafür, welche Methoden in einem bestimmten Umfeld zu nutzen sind. Dieser Beitrag stellt eine Klassifikation der Evaluationsmethoden vor mit dem Ziel, deren Auswahl zu vereinfachen. Die Klassifikation basiert auf Angaben zu zwei grundlegenden Fragestellungen: Zweck und Ziel der Evaluation sowie den jeweiligen Rahmenbedingungen für das Lernen. Es besteht die Hoffnung, daß die Klassifikation hilft, die Qualität computer-gestützten Lehrens dadurch zu verbessern, daß ein praktikables und zielgerichtetes Verfahren zur Auswahl geeigneter Evaluationsmethoden für intelligente tutorielle Systeme bereitgestellt wird.

1 Introduction

Intelligent tutoring systems are increasingly being employed in education ([18], [21]) and as a consequence the need for careful and systematic scrutiny of these systems has become an important issue. The literature provides a number of evaluation methods that can be used to evaluate intelligent tutoring systems, but given the diversity of these methods, it is quite difficult for an evaluator, normally an educator, to judge which method is appropriate for a particular purpose. [39] argued that there are no guidelines available for use of evaluation methods and most of the existing evaluations are empirical endeavours. Although such empirical experiments are necessary for further research, they do not allow a generalisation of the approach, and may fail to consider all possible independent factors leading to inconclusive results and debatable conclusions.

This paper proposes a classification of evaluation methods based on two primary questions an evaluator needs to ask before conducting the evaluation:

i) What is being evaluated: the whole system or just a component?

ii) Is it possible to systematically manipulate variables in the evaluation, and how many users are available for the purpose of evaluation?

The evaluation methods are classified along two dimensions, each relating to one of the above questions. The first dimension focuses on the the *degree of evaluation* covered by the evaluation method. [17] pointed out that if a method solely concentrates on testing a component of a system, it can be considered suitable for *internal evaluation*. If the method evaluates whole system, it is suitable for *external evaluation*.

The second dimension is concerned with the local feasibility of using a particular evaluation method. It distinguishes between a method *establishing a causal association* from a controlled investigation i. e. *experimental research* and the *accumulation of a large amount of data* about a particular aspect of the system i. e. *exploratory research*. *Experimental research* requires experiments varying systematically the independent variable(s) while measuring the dependent variable(s) and ensuring random assignment of participants to conditions and require statistically significant groups. *Exploratory research* includes in-depth study of the system in a natural context using multiple sources of data, usually where sample size is small and the area is poorly understood.

The paper starts by discussing the great necessity for the evaluation of intelligent tutoring systems and then reviews a number of evaluation methods for intelligent tutoring systems suggested by various researchers. This is followed by the description of proposed classification of evaluation methods. The paper concludes by giving an example of the evaluation methods selection using the proposed classification.

2 Ensuring the effectiveness of ITSs

[23] pointed out that evaluation of ITSs has not been taken seriously. [20] noted that most ITS researchers have concerned themselves only with envisioning the potential of ITSs and investigating the implementation issues involved in constructing actual components and systems, and have paid little attention to the process of evaluation. [33] commented that despite the decade long existence of ITSs, the degree to which they have been successful is equivocal solely due to the lack of proper evaluation.

Though the evaluation of intelligent tutoring systems is a costly and time consuming affair, [17] argued that it pays off by helping to answer two questions that are central to cognitive science, artificial intelligence, and evaluation:

i) What is the educational impact of an ITS on students?

ii) What is the relationship between the architecture of an ITS and its behaviour?

Another advantage of evaluation is that it provides an opportunity to learn from the mistakes and is capable of improving the life-span of ITSs as well as their usability. [17] acknowledged that by evaluating PROUST, they learned a lot about how novices learn to program, how to teach programming, and how to build ITSs to actually do the teaching. [33] also expressed the same enthusiasm for evaluation by accepting that they found the results of careful experimental design always informative.

Furthermore, the benefit of carrying out ITS evaluation is to focus the attention away from short-term delivery and open up a dialogue about issues of appropriateness, usability and quality in system design. Evaluation can provide benefits of shared experiences and help in avoiding continuously reinventing the wheel. Researchers have felt the need to develop a more systematic approach to the evaluation of ITS which means ensuring availability of evaluation methods and adequate quality control in the evaluation process. [38] underlined this point by forecasting that "*in a future where [AI-ED systems] may be widely available to schools and training institutes, evaluations will shape what and how people learn, and what they become able to do*".

Though many researchers have attempted to formalise the evaluation procedures of ITSs from different points of view, these formal methods vary from each other as much as the design and development methodologies of various ITSs vary. We take these formal evaluation methods as the basis of our research. The following section provides summary of these methods. The list of methods is not meant to be exhaustive, but covers a majority of methods discussed in the literature for the evaluation of ITSs.

3 Review of evaluation methods

The evaluation methods for ITSs suggested by various researchers provide a large variety of factors to be considered and the evaluation seems not to be a trivial task, as [20] pointed out: "*There are few agreed upon standards within the ITS community to guide investigators who wish to evaluate systems. However, other fields have developed evaluation methods which may be applicable to ITSs.*" The methods discussed in this section are adapted by various ITS researchers from expert systems development, computer-based instruction, education, evaluation, computer science, engineering and psychology disciplines.

Methods for experimental research

The methods in this category require systematically varying the independent variable(s) while measuring dependent variable(s) and ensuring randomly assignment of participants to conditions and require statistically significant groups. As reasoned below, the first four methods are suitable for internal evaluation due to their focus on testing components of the systems rather than overall effectiveness. Methods 6 and 7 focus on overall effectiveness and therefore are good for external evaluation. Method 5 can be used for both component testing or overall effectiveness and hence is suitable for both internal and external evaluations.

1. Proof of correctness is a check whether the system fulfils the desired requirements or goals or whether there is a correspondence between its structure and behaviour and its specifications. The method evaluates the internal components of the systems in rather hypothesis testing fashion and therefore is suitable for internal and experimental nature of evaluation. [20] argued against this method in the case of ITSs due to the inherent AI nature of ITSs which deals with analytically intractable problems represented as incompletely specified functions [28].

2. Additive experimental design comparisons have been proposed to evaluate the impact of *large* ITS components that can be experimentally modified or withheld [24]. Thus they belong to experimental research and have limited suitability only for internal evaluations. These comparisons have been used for ITS component evaluation by [35] for competing tutoring approaches in Pixie tutor and by [3] for the effect of strategy workbooks in West tutor. This approach requires a large number of students and is not cost-effective if the ITS component is minor. The advantage of additive design includes possibility of individual aspects of the ITS to be manipulated in order to directly assess their impact and importance.

3. The **diagnostic accuracy** of an ITS can be estimated through procedures that address the quality of micro-theories used by the system. [16] contended that the method of diagnostic accuracy has gained more prominence for internal evaluation of ITSs because subsequent pedagogical interactions are dependent on the correct recognition and interpretation of student errors. The method provides a controlled examination of the student model across conditions and therefore belongs to experimental research. This method has been used to evaluate POSIT ITS by [25].

4. The immediate impact of **feedback/ instruction quality** on the student can be experimentally estimated through lag sequential analysis procedures which are suitable for internal evaluation. One way to measure this is to calculate a lag sequential probability, that is the ratio of actions in a specific category to the total number of actions that occur within a specific frame following a target action. The advantage of such a measure is that a calculation can be made for either a different level of detail in ITS feedback or even for different types of feedback. The ALM tutor ([26]) was evaluated for its feedback quality using lag sequential prabability.

5. In **sensitivity analysis**, an examination is made of how varying information to the component or system would result in diverse responses. This experimental approach is equally suitable for components and whole ITS evaluations and hence can be used for both internal and external evaluations. [20] argued that this method is relevant for ITSs since individualised instruction is a primary objective in intelligent tutoring.

6. The **experimental research** enables the researchers to obtain a relationship between a set of interventions and the outcome and as a consequence it is particularly suited to examining the effects of teaching. [20] pointed out a variety of experimental research designs such as single group designs, control group designs, and quasi-experimental designs. Experimental research has been used in ITS evaluation by [14] for Sherlock, [19] for VCR tutor and by [13] for Byzantium ITTs. Experimental research is more appropriate for external evaluations since it is likely to provide overall conclusions rather than acquisition of information [20].

7. [15] examined **product evaluation** as broad-based ITS evaluation studies to identify promising evaluation procedures that could justify continued ITS development. After analysing past ITS evaluation studies (for example, [3], [35]), three guidelines were specified:

i) the instructional effectiveness of ITS applications, human tutors, and traditional methods need to be compared on the basis of performance data;

ii) the instructional effectiveness of only extensive ITS applications should be evaluated; and

iii)large groups of subjects are required to precisely estimate ITS effectiveness.

Since the method manipulates three classic learning conditions, it belongs to experimental research.

Methods for exploratory research

Methods in this category involve an in-depth study of the system in a natural context using multiple sources of data, usually where sample size is small and the area is poorly understood. Similar to the reasoning given in experimental research methods section, methods 8 to 12 in this category are suitable for internal evaluation due to their focus on part evaluation rather than whole system evaluation. Methods 15 to 20 are suitable for external evaluation and methods 13 and 14 can be used for both.

8. An altogether different approach is to make use of an **expert inspection** to assess whether a program meets an explicit standard level of performance. This type of evaluation is often used in the development of knowledge-based systems and the most well-known example is the Turing test which compares human and computer behaviour [27]. Such inspection is useful for evaluating specific ITS components making the approach suitable for internal evaluations. This method can be classified as exploratory since no controlled hypothesis testing takes place.

9. The **level of agreement** between subject matter experts is related to knowledge base varification and can be estimated correlationally and is suitable for internal evaluations. This approach uses subject experts or manuals and correlational methods to estimate the consistency of knowledge and beliefs across experts. Therefore the method falls into exploratory research. The knowledge base of Media Selection Expert System (MSES) was verified using agreement of media specialists by [4].

10. Wizard of Oz experiments use a human to simulate the behaviour of a proposed system, so that one can test aspects of a program design before actually implementing it [36]. This method is more favourable for internal evaluation than external ones. [30] used this method for an intelligent help system of a text editor. The difficulty in predicting the way the human will behave makes it impossible to decide in advance, the way the parameters should be observed. This method therefore belongs to exploratory research.

11. Performance metrics are methods which use quantitative results to explore individual factors or features of an ITS and thus fall into the category of internal evaluations. These are important in exploratory settings when data pools are too small to allow statistically significant conclusions. But such methods of measuring system performance include false positive and false negative results. These methods encompass three sub-categories [22]: *diagnostic accuracy*, where measures of the success of ITS components are assessed by recording 'hit and miss rates' of students; *feature usage*, which assesses how users utilise various features of tutor; and *cognitive change*, which can describe differences across populations or time periods, as well as absolute metrics. [34] and [12] evaluated various ITS components using performance metrics methods.

12. [22] pointed to a category of **internal evaluation** methods which study the relationships between an ITS architecture and its behaviour, and often involve in-depth exploratory analysis of program traces and data structures. This method includes following sub-categories: *knowledge level analysis* to evaluate knowledge base, for example, [5] evaluated MYCIN to make NEOMYCIN for intelligent tutoring; *process analysis* for evaluation of algorithms, for example, [17] analysed the limitations of PROUST system; and *ablation & substitution experiments* ([6]) to observe system performance by replacing parts of the system with more primitive versions, for example, [7] analysed LISP tutor without its feedback capabilities.

13. A similar method is **criterion-based** evaluation, where the requirements and specifications are drawn up almost as a checklist which are then used in an exploratory way to compare the systems inadequacies. This method can be used to examine either the components of the ITSs or the whole system and hence is suitable for both internal and external evaluations. This method has been used by [9] for ITS evaluation.

14. The **pilot testing** method is used to check the system design for unanticipated outcomes of using the system with users. The method is in essence an exploratory tool for identifying problems in the system. Three types of pilot tests are identified ([10]) as providing opportunities to learn about different parts of the development process both at component and whole system level. *One-to-one* pilot tests are useful in the early stages of development. They have been used by [29] to evaluate 'Guidon-watch' knowledge-base interface. *Small group testing* is conducted later in the development with a representative group of students once the system has begun to stabilise and finally *field testing* takes place towards the end when completion is only a short period away.

15. Another approach to ITS evaluation is **certification**, based on methods currently applied in identifying competent human teachers. Such a procedure has the benefits of an authoritative endorsement of the correctness of the ITS and due to this reason it falls under exploratory research. The suitability of the approach for overall program makes it adequate for external evaluations. [20] favoured this method as a way of obtaining feedback on the strengths and weaknesses of ITSs in formative evaluation and of obtaining adequacy ratings in summative evaluation.

16. The opinions of experts or a large number of (potential) users of a system, known as **outside assessment**, can be significant if there is agreement. [22] argued that there are two types of such assessments: *on-site expert evaluation* and *panel of experts*. The former is the normal way of getting experts to observe and assess the behaviour of system, for example several teachers rating an ITS effectiveness with some students using it. The latter relies on a selected group of experts being questioned by convening them or through correspondence. Since the method operates on overall ITS and is based upon an exploration of users' opinions, it is suitable for exploratory external evaluations.

17. Existence proofs method bases its conclusions on the successful implementation of a system or application of a method to propose a new system architecture or design method. Since the method applies on whole system, it is suitable for external evaluations. This method is only powerful if the researcher describes the goals and assumptions of the study, discusses design trade-offs, and documents surprises and failures. It is useful in exploratory research in beginning to identify key issues and providing a baseline from which further studies can be designed [22]. This method has been used by [2] and [1] showed by different means that various ITSs were possible.

18. The intention of **observation and the qualitative classification of phenomena is to identify classes of phenomena**, patterns, and trends in the interaction of people with instructional systems. This data is gathered for exploratory research primarily through observation, anywhere from natural to contrived situations as part of external evaluations. [22] stated that "*The observation can be of novices (learners) or experts (teachers), and can be aimed at identifying behavioural phenomena or at inferring cognitive phenomena*".

19. The approach of focusing on the collection of data by limiting or organising the responses of the subject under exploratory research is collectively known as **structured tasks and the quantitative classification of phenomena**. Here quantitative data is gathered in contrived or structured situations in external evaluations, for example by interviews, questionnaires and surveys. The benefit of this approach is that more voluminous and precise data can be obtained by using more structured tasks. The method has been used by [31] in a study of Anderson's LISP tutor.

20. Comparison studies are a general class of methods which note similarities and differences between the behaviour or design of an ITS with a standard or even another system. These methods are suitable for external evaluations and apply to overall system performance in exploratory research. [22] suggested three methods to achieve such comparisons: judging system performance against a well-known successful case (i.e. gold standard), for example, [9] evaluated Highway code tutor against [32]'s definition of "intelligence"; arguing at a theoretical level for the generality and extendibility of a new approach to other systems (i.e. theoretical corroboration), for example, [8] compared PRO-TEG system with six other ITSs using a three dimensional ITS classification system; and, simulating the behaviour of another system (i.e. empirical corroboration and duplication).

4 Proposed classification of evaluation methods

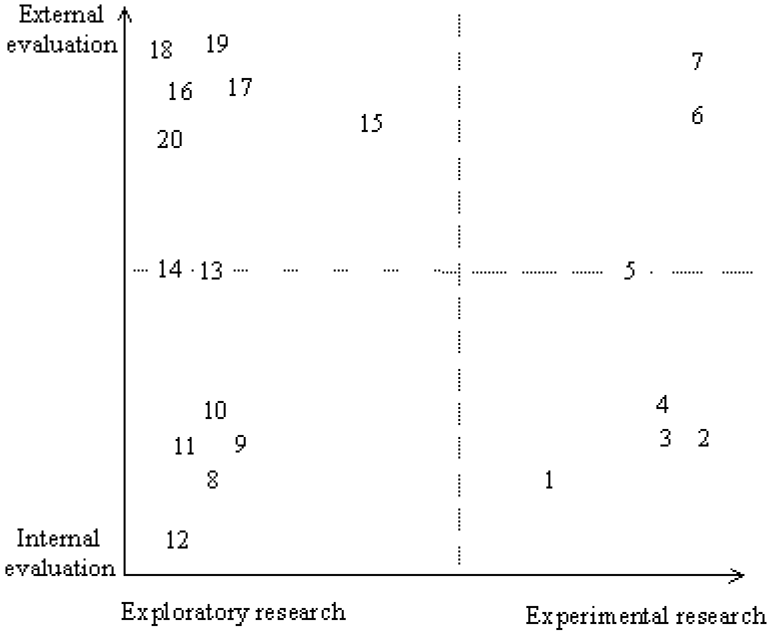
Given the variety of evaluation methods, it is difficult to decide which one is appropriate in a particular context. From an examination of the evaluation requirements two important questions emerge which confront the educator who is going to employ the ITS and hence evaluating it:

(i) What is being evaluated - the whole system or simply a part of the system?

(ii) Are there enough students and other suitable conditions available to conduct evaluation based on an experimental design?

These two questions can be used to classify various methods, so that a method could be differentiated from a number of others on a scale between external evaluation (considering the whole system) and internal evaluation (testing a component of the system). In addition, a method could be classified along a dimension consisting of exploratory research versus experimental research. In the case of exploratory research such methods should be applicable where samples are small and the area is poorly understood. Evaluation methods which are categorised as experimental are likely to involve manipulation of variables and require statistically significant groups. Based on the illustration of each method in the section 3 of this paper, figure 1 classifies the methods relative to these two dimensions.

The methods falling towards lower left part of the chart are suitable for evaluation of ITS components and do not require large sample sizes or rigorous statistical studies. Methods which require control group studies and statistical procedures for component evaluation are grouped towards the lower right part of the chart. Upper part of the chart shows methods which evaluate overall system performance. The methods at the right side require statistical studies whereas the methods at left side do not require such studies.



Key to evaluation methods:

1. Proof of Correctness
2. Additive experimental design
3. Diagnostic accuracy
4. Feedback/instruction quality
5. Sensitivity Analysis
6. Experimental research
7. Product evaluation
8. Expert knowledge
9. Level of agreement
10. Wizard of Oz experiment
11. Performance metrics
12. Internal evaluation
13. Criterion-based
14. Pilot testing
15. Certification
16. Outside assessment
17. Existence proofs
18. Observation & qualitative classification
19. Structured tasks & quantitative classification
20. Comparison studies

Figure 1: Classification chart of evaluation methods

The following section gives an example of using the classification chart for selection of methods appropriate to the local requirements and conditions at hand.

5 Selecting evaluation methods by classification chart – An example

An evaluation was carried out on a fully operational intelligent tutoring system [37], aimed to teach marginal costing subject in cost engineering/ accounting domain, to assess its value and effectiveness with target users (i.e. students). The evaluation was conducted at three universities in United Kingdom. The results of evaluation are compiled in [11].

The research design for evaluation was selected with reference to the aims of the study, that is firstly a comparison of how effective computer aided instruction was in relation to conventional teaching and secondly a determination of the features significant in assessing a teaching system. The classification chart was used to select the methods used for evaluation. At first the two evaluation requirements were examined:

(i) What is being evaluated - the whole system or simply a part of the system?

(ii) Are there enough students and other suitable conditions available to conduct evaluation based on an experimental design?

Since the study aimed to evaluate the whole system and not just a part of the system it was classified as an external evaluation. However, the other question provided a mixed reply. The comparison of instruction techniques demanded testing of hypothesis to evaluate the question: which type of instruction is “better”. This led to the conclusion that an experimental design should be used. On the other hand the requirement to collect a multifaceted assessment from the users of the computer aided instruction suggested an exploratory evaluation. As a consequence two evaluation methods were chosen, one each from exploratory and experimental techniques but both suitable for external evaluation. A structured questionnaire was used simulating *Structured tasks and the quantitative classification of phenomena* (labelled as 17 in the chart) to collect information from users for multifaceted assessment and then a before-after two-group experiment was used for *Product Evaluation* (labelled as 12 in the chart) to determine the differences between two instruction methods.

6 Conclusion

The usefulness of the proposed classification of evaluation methods for ITSs is in the fact that it enables the evaluators to select the most appropriate evaluation methods for their investigation. The evaluators have to decide whether the investigation will primarily involve the whole system or just a few components. Then they have to judge whether it is possible to conduct the evaluation in an experimental style. At this stage, it should be clear from which region of the chart the method has to be picked. The final decision is dependent on practical factors faced by the evaluator and the quality of information being sought, and ultimately the chosen method is characteristic of those individual circumstances.

While the proposed classification provides a simple yet robust way to select evaluation methods, the classification requires a fully developed ITS available in hand. The ITS field in fact, has very few fully developed products in market. Most of the systems are still at their prototyping stages, confined in the research laboratories. To get most out of the research, future work is planned to add another dimension of formative vs. summative evaluation to the classification chart. This dimension will facilitate the educators to formatively evaluate the prototypes for adequacy to their curriculum and cohorts while keeping the classification open for summative evaluations of any fully developed products.

Acknowledgement

The research presented in this paper is partially supported by Science and Engineering Research Council (UK) studentship awarded to the first author.

References

[1] Buchanan B.: Artificial intelligence as an experimental science. Stanford University Knowledge Systems Lab Technical Report KSL 87-03, 1987.

[2] Burke R. & Funaro G. M.: Case-based environments for learning. Working Notes of the AAAI Spring Symposium on Knowledge-Based Environments for Learning and Teaching, March, 1990, Stanford, CA, 63-67.

[3] Center for the Study of Evaluation: Intelligent computer aided instruction (ICAI): Formative evaluation of two systems (ARI RN 86-29). Alexandria, VA: U. S. Army Research Institute, 1986.

[4] Chao P. C. & Legree P. J.: The Media Selection Expert System knowledge base verification and performance evaluation analyses. ARI Technical Report, U. S. Army Research Institute, VA, 1991.

[5] Clancey W. J.: Tutoring rules for guiding a case-method dialogue. In: Sleeman D. & Brown J. S. (eds.) Intelligent Tutoring Systems, New York, 1988: Academic Press, 201-225.

[6] Cohen P. & Howe A.: How evaluation guides AI research. AI Magazine, Winter (1988).

[7] Corbett A. T. & Anderson J. R.: The effect of feedback control on learning to program with the LISP tutor. Proceedings of the Twelfth Annual Conference of the Cognitive Science Society, July, 1990, Cambridge, MA, 796-806.

[8] Dillenbourg P.: A model of knowledge acquisition by intelligent tutoring systems. Proceedings of ITS-88, 1988, Montreal, Canada, 145-153.

[9] Ford L.: The appraisal of an ICAI system. In: Self J. (ed.) Artificial Intelligence and Human Learning: Intelligent Computer-aided Instruction, London, 1988: Chapman & Hall.

[10] Gagne R. M., Briggs L. J. & Wager W. W.: Principles of Instructional Design, New York, 1988: Holt, Rinehart and Winston.

[11] Iqbal A. M.: Evaluation of computer aided instruction: Assessing the value and effectiveness of operational systems. PhD thesis submitted for assessment, University of London, UK, 1997.

[12] Kimball R.: A self-improving tutor for symbolic integration. In: Sleeman D. & Brown J. S. (eds.) Intelligent Tutoring Systems, New York, 1988: Academic Press, 283-307.

[13] Kinshuk: Computer aided learning for entry level Accountancy students. PhD Thesis, De Montfort University, England, July, 1996.

[14] Lajoie S. P. & Lesgold A. M.: The SHERLOCK experience: An evaluation of a computer-based supported practice environment for electronics trouble-shooting training. Proceedings of the International Conference for Cognitive Science for the Development of Organizations, May 2-4, 1991, Montreal, 56-62.

[15] Legree P. J. & Gillis P. D.: Product effectiveness evaluation criteria for intelligent tutoring systems. Journal of Computer-Based Instruction, 18 (2) (1991), 57-62.

[16] Legree P. J., Gillis P. D. & Orey M. A.: The quantitative evaluation of intelligent tutoring system applications: Product and process criteria. Journal of Artificial Intelligence and Education, 4 (2/3) (1993), p209-226.

[17] Littman D. & Soloway E.: Evaluating ITSs: The cognitive science perspective. In: Polson M. C. & Richardson J. J. (eds.) Foundations of Intelligent Tutoring Systems, New Jersey, 1988: Lawrence Erlbaum Associates, 209-242.

[18] Major N., Ainsworth S. & Wood D.: REDEEM: Exploiting sumbiosis between psychology and authoring environments. International Journal of Artificial Intelligence in Education, to appear.

[19] Mark M. A. & Greer J. E.: The VCR Tutor: Evaluating instructional effectiveness. Proceedings of the 13th Annual Conference of the Cognitive Society, 1991, 564-569.

[20] Mark M. A. & Greer J. E.: Evaluation methodologies for intelligent tutoring systems. Journal of Artificial Intelligence and Education, 4 (2/3) (1993), 129-153.

[21] Mizoguchi R., Sinitsa K. & Ikeda M.: Task ontology design for intelligent educational/training systems. Position paper for ITS'96 Workshop on Architectures and Methods for Designing Cost-Effective and Reusable ITSs, June 10th 1996, Montreal.

[22] Murray T.: Formative qualitative evaluation for `exploratory' ITS research. Journal of Artificial Intelligence and Education, 4 (2/3) (1993), 179-207.

[23] Nwana H. S.: Mathematical Intelligent Learning Environments, Oxford, 1993: Intellect Books.

[24] O'Neil H. & Baker E.: Issues in intelligent computer aided instruction: Evaluation and measurement. In: Conoley J. C. (ed.) The Computer as Adjunct to the Decision Making Process, Hillsdale, NJ, 1987: Lea/Wiley.

[25] Orey M. A. & Burton J. K.: POSIT: Process oriented subtraction – interface for tutoring. Journal of Artificial Intelligence in Education, 1(2) (1989/90), 77-104.

[26] Orey M. A., Park J. S., Chanlin L. J., Jih H., Gillis P. D., Legree P. J. & Sanders M. G.: High bandwidth diagnosis within the framework of a microcomputer-based intelligent tutoring system. Journal of Artificial Intelligence in Education, 3(1) (1992), 63-80.

[27] Parry J. D. & Hofmeister A. M.: The development and validation of an expert system for special educators. Learning Disability Quarterly, 9(2) (1986), 124-132.

[28] Partridge D.: Artificial Intelligence: Applications in the future of software engineering, New York, 1986: Ellis Horwood.

[29] Richer M. H. & Clancey W. J.: Guidon-watch: A graphic interface for viewing a knowledge-based system. In: Lawler R. W. & Yazdani M. (eds.) Artificial Intelligence and Education: Volume one, Learning Environments and Tutoring Systems, Norwood, NJ, 1987: Ablex.

[30] Sandberg J., Breuker J. & Winkels R.: Research on HELP-Systems: Empirical study and model construction. Submitted for ECAI, Munchen, 1988.

[31] Schofield J. & Verban D.: Barriers and incentives to computer usage in teaching. LRDC Technical Report No. 1, 1988.

[32] Self J.: Intelligent computer assisted instruction. Unpublished paper presented at the ICAI Spring Seminar, 1985, Cambridge.

[33] Shute V. J. & Regian J. W.: Principles for evaluating intelligent tutoring systems. Journal of Artificial Intelligence and Education, 4 (2/3) (1993), 245-271.

[34] Sleeman D.: Assessing aspects of competence in basic algebra. In: Sleeman D. & Brown J. S. (eds.) Intelligent Tutoring Systems, New York, 1988: Academic Press, 185-199.

[35] Sleeman D., Kelly A. E., Martinak R., Ward R. D. & Moore J. L.: Studies in the diagnosis and remediation of high school algebra students. Cognitive Science, 13 (1989), 551-568.

[36] Twidale M.: Redressing the balance: The advantages of informal evaluation techniques for intelligent learning environments. Journal of Artificial Intelligence and Education, 4 (2/3) (1993), 155-178.

[37] Wilkinson-Riddle G. J. & Patel A.: "Human Tutor" Emulation in Teaching Numerical Business Subjects - a Software Breakthrough. Proceedings of the tenth International Conference on Technology and Evaluation, March 21-24, 1993, 174-176.

[38] Winne P. H.: A landscape of issues in evaluating adaptive learning systems. Journal of Artificial Intelligence and Education, 4 (4) (1993), 309-332.

[39] Wu A. K. W.: On the formal evaluation of learning systems. Lecture Notes in Computer Science, 1086 (1996), 324-332.

Authors' addresses

Mr. Arif Iqbal
42, New Rowley Road
Dudley, West Midlands DY2 8AS
United Kingdom
Tel/Fax: (+44) 1384 459 129

Prof. Dr. R. Oppermann
GMD-FIT, German National Research Center for Information Technology
Schloss Birlinghoven, 53754 St. Augustin
Germany
Email: oppermann@gmd.de

Mr. A Patel
CAL Research & Software Engineering Centre,
Bosworth House, De Montfort University
The Gateway, Leicester LE1 9BH
United Kingdom
Email: apatel@dmu.ac.uk

Dr. Kinshuk
GMD-FIT, German National Research Center for Information Technology
Schloss Birlinghoven, 53754 St. Augustin
Germany
Email: kinshuk@ieee.org