

The Evaluation of Adaptive Technology Enhanced Learning Systems

Catherine Mulwa, Seamus Lawless, Mary Sharp, Vincent Wade

Centre for Next Generation Localisation
School of Computer Science and Statistics,
Trinity College Dublin,

{mulwac, seamus.lawless, mary.sharp, vincent.wade}@scss.tcd.ie,

Abstract

Adaptive technology enhanced learning has attracted significant interest with the promise of supporting individual learning tailored to the unique circumstances, preferences, and prior knowledge of a learner. Evaluation of the overall performance of such adaptive TEL systems is a major challenge; as such systems react differently for each individual user and context of use. Evaluation of such systems has become a significant but very complex area of research in itself since depending on the aspect of adaptivity and personalisation that needs to be evaluated (quality of the user modelling, performance of different adaptation approaches, knowledge gain from using the personalised system or overall end user experience), several evaluation techniques need to be combined and executed differently. This paper proposes a hybrid recommendation service for recommending appropriate evaluation techniques (approach, methods, metrics and criteria). It also discusses evaluation challenges and presents analysed results of a survey on evaluations of adaptive systems.

Keywords: Evaluation Challenges, Adaptive Technology Enhanced Learning, Recommendation Service

1. Introduction

Adaptive technology enhanced learning (TEL) has attracted significant interest with the promise of supporting individual learning tailored to the unique circumstances, preferences, and prior knowledge of a learner. It is important not only to evaluate the adaptive TEL system but also ensure the correct evaluation techniques are used and to value the integration within the content where the learning will take place. A major challenge faced by evaluators of adaptive TEL systems is the difficulty in choosing the right evaluation approach and technique to use. The evaluation of adaptive TEL systems is significant and a challenging difficult task (Lawless et al., 2010, Tintarev and Masthoff, 2009). One major problem is the understanding of the adaptation mechanism of the system; what is improved by the adaptation and what might have been the situation if a different kind of adaptation had occurred. It is difficult in defining the effectiveness of adaptation. When users work with an adaptive system, it is very difficult in principle to demonstrate what "might have been" or what impact the systems' adaptive processes actually made on the end-user. Several researchers have emphasized the difficulties caused by the complexity of such systems and the usability issues raised by end users (Missier Del and Ricci, 2003, Lavie et al., 2005, Weibelzahl and Weber, 2002, Markham et al., 2003). Major challenges include usability issues such as: i) usability goals correspond to several desirable properties of interactive systems, ii) predictability, transparency, controllability and unobtrusiveness correspond to general usability principles, iii) maintenance of privacy and breadth of experience are relevant to adaptive and personalised e-learning systems, iv) the column typical properties lists down existing examples of frequently encountered properties of these systems. Each has the potential of causing difficulties with respect to one or more of the usability goals, v) the preventive measures aim is to ensure that a property is not present in such a manner that it would cause problems and vi) compensatory measures goal is to ensure in some other way that the goals and objectives are achieved despite the threats created by the properties challenges. Furthermore evaluation of these systems is a crucial and significant stage in their development (Jameson, 2009). These systems require some kind of evaluation due to their inherent usability problems at the interface and for correctness of adaptive solutions.

This research investigates current evaluation challenges encountered by evaluators of adaptive systems and tries to address the question of: "What are the evaluation techniques used (and benefits of) evaluating adaptive TEL systems: Can these techniques be appropriately combined and applied to evaluate such systems". Evaluation techniques are concrete methods to carry out the validation of the system. The main contribution of this paper is a review of evaluation procedures for adaptive TEL systems. It also proposes a hybrid recommendation service for evaluating adaptive TEL systems developed from 2000 to 2012. We use recommendation technology to enhance the appropriateness of suggestions of evaluations approaches of adaptive TEL systems. In particular the multi attribute relationships which need to be traversed by humans to work out what are the most appropriate

evaluation procedures (i.e. evaluation approaches, methods/techniques, metrics and criteria) are not easily navigated using typical database techniques.

This paper is structured as follows. Section 1 presents current challenges encountered by evaluators of adaptive TEL systems. Section 2 presents a survey of evaluation of adaptive TEL systems and talks about impact of personalisation and personalised learning. Section 3 proposes a recommendation service developed for evaluating adaptive systems. Finally section 4 concludes the paper and recommends future work.

2 A Review on the Evaluation of Adaptive TEL Systems

Technology-enhanced learning environments (TELE) increasingly offer possibilities for adapting and personalising learning activities and experiences. When technologies are intergraded into a single environment or platform to accomplish the goal of enhancing student learning via adaptation, this is called “adaptive learning environments (e.g. adaptive hypermedia, adaptive educational hypermedia, collaborative learning and simulation immerse environment)”. These environments provide error feedback that is tailored to the learner or enable the learner to customize the learning environment to fit user’s interests. Most relevant literature on adaptive learning is focused on adaptivity; by adaptivity we mean the possibility for learners to personalise the course materials themselves (Burgos et al., 2007). In order to address the challenges reported in section 1, we conducted a study on evaluations of adaptive systems and also the academic impact of personalisation. Personalised learning focuses on improving the efficiency and cost-effectiveness of learning for individuals and organizations, independent of time place and pace through the development of open systems and services.

2.1 A Survey of the Evaluation of Adaptive TEL Systems

Over the last three years, we have been reviewing evaluations techniques for adaptive systems developed from 2000 to date (Mulwa et al., 2011b, Mulwa et al., 2011a, Mulwa C. et al., 2010). Recently we conducted a study on the evaluation of adaptive TEL systems developed from 2000 to 2011. Table 1 presents the survey question. If the reader wishes to fill in the survey is available online (<http://surveymonkey.com/s/Q2DSDF8>). Over 100 domain experts from the User Modelling, Adaptation and Personalisation (UMAP), Adaptive Hypermedia (AH) and Adaptive Recommender communities participated. The aim was to investigate current evaluation procedures of adaptive systems. The analysed data was used to create an educational dataset which we are currently using to populate the recommendation service proposed in section 3.

Table 1: Survey Questions

No	Questionnaire
1.	Have You Developed an Adaptive System in the Past (from 2000 to 2011)? (i.e. An Adaptive System Refers to a System which Tailors its Output, using Implicit Inferences based on Interaction with the User) Yes [] No []
2.	If You Answered Yes to this Question, Please Provide: i) Name of Adaptive System, ii) Year the System was Developed, iii) Other Details
3.	If You Have Developed an Adaptive System(S), what was improved by Adaptivity?
4.	What is the Variation Type of the Adaptive System You have Developed?
5.	Please Tick the Meta Data Models Your System Uses
6.	If You Conducted a Whole-System Evaluation, What Evaluation Methods did you use?
7.	If you conducted a whole evaluation, what criteria did you use?
8.	If You Conducted Evaluations of Specific Metadata Models of Adaptive System, What Evaluation Methods did you use? (For each model evaluated, please indicate which evaluation methods and criteria you used)
9.	During this Evaluation (Conducted in Question 6 and 7 above), What Metrics did You Use to Measure Performance against these criteria?
10.	A Web-Based Evaluation Framework for End User Experience in Adaptive Systems (EFEx) Usefulness of EFEx Framework: i) Repository for User-Centred and Layered Evaluations of Adaptive Systems. ii) Recommendations on How to Evaluate an Existing Adaptive System or a New Adaptive Systems. iii) Recommendations on How to Evaluate Metadata Models of Adaptive Systems, iv) User-Centred Evaluation Methodology for Adaptive Systems, v) Taxonomy of Evaluations of Adaptive Systems and v) Information Translated into Users' Language of Choice ----- Which of the following features of EFEx Framework would you find (consider) useful?

In Q1, we wanted to gauge the interest of the user in adaptive systems. The aim of Q2 was to identify how many adaptive systems were developed from 2000 to 2011, so that we can check whether adaptivity had any impact to learners in adaptive TELE. Q3 aimed at identifying the facets of adaptivity and impact of adaptivity to learners.

In Q4 and Q5 aimed at identifying which of the reported systems belonged to TEL variation type (category) and what metadata models were used. The most important questions were Q6, 7, 8 and 9; whose aim was to investigate how such systems were evaluated and the techniques (approaches, methods, metrics and criteria) used.

This research has resulted in the implementation of an evaluation framework for supporting expert and novice evaluators of adaptive systems (EFEx). The aim of Q10 was to find out which components of the framework would be valuable to developers of adaptive systems. EFEx is divided into four major components. One of these components is the proposed recommendation service in section 3. Following is the analysed results of this study:

Figure 1: Total Number of Participants

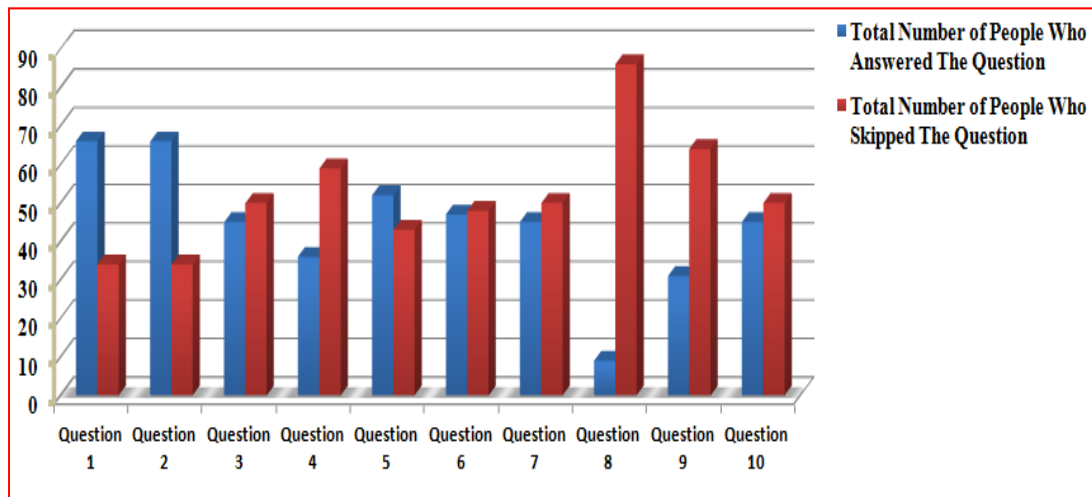


Figure 2: A summary of survey analysis

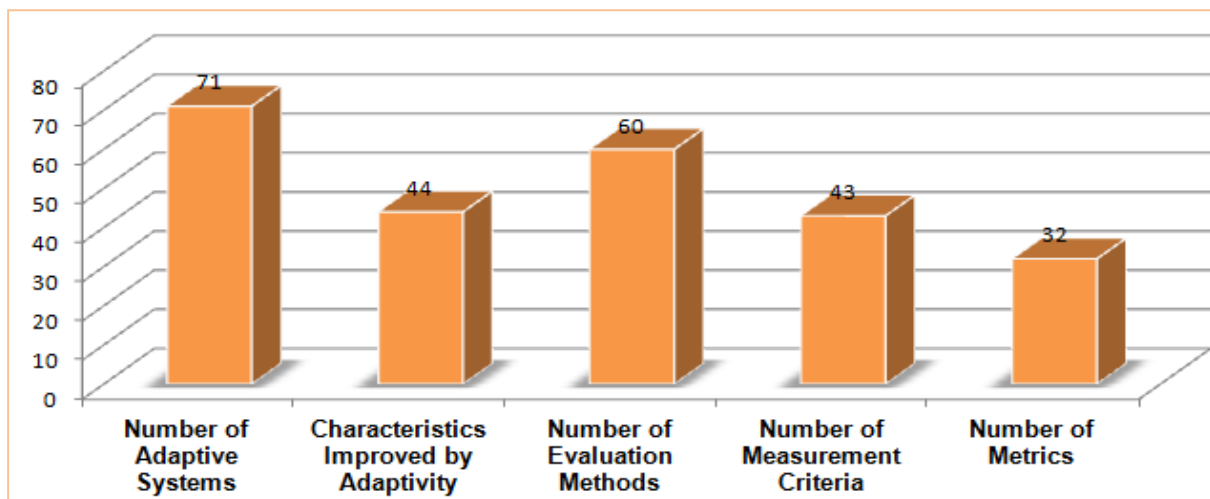
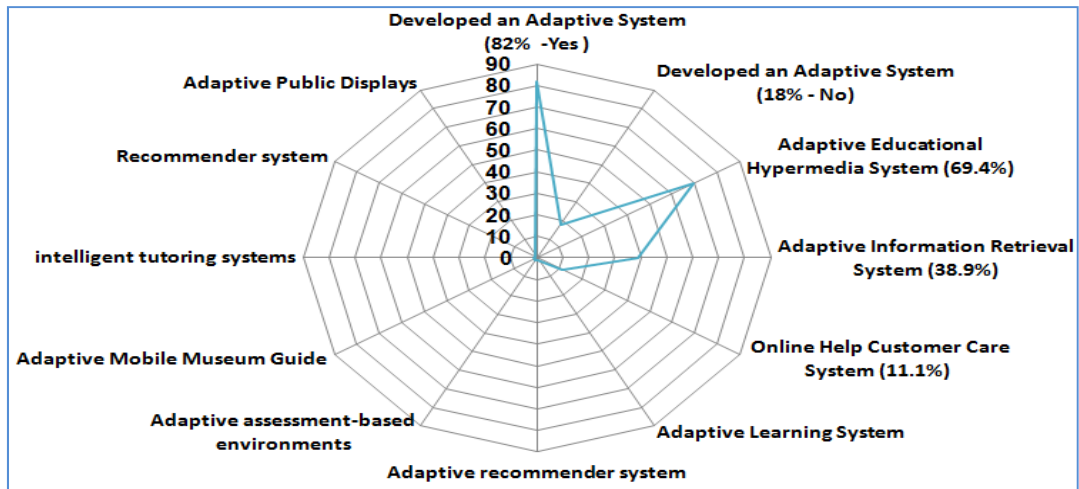


Figure 2 presents a summary of analysis of Q1 to 9. We characterized, structured and interlinked the results to form a list of adaptive systems, evaluation methods, metrics and measurement criteria (Figure 2). Over 77 adaptive systems were mentioned in the study. These systems belonged to several variation types (Figure 3) of which Majority were adaptive educational hypermedia systems with a 69.4% response count.

Figure 3: Variation Types of Adaptive Systems

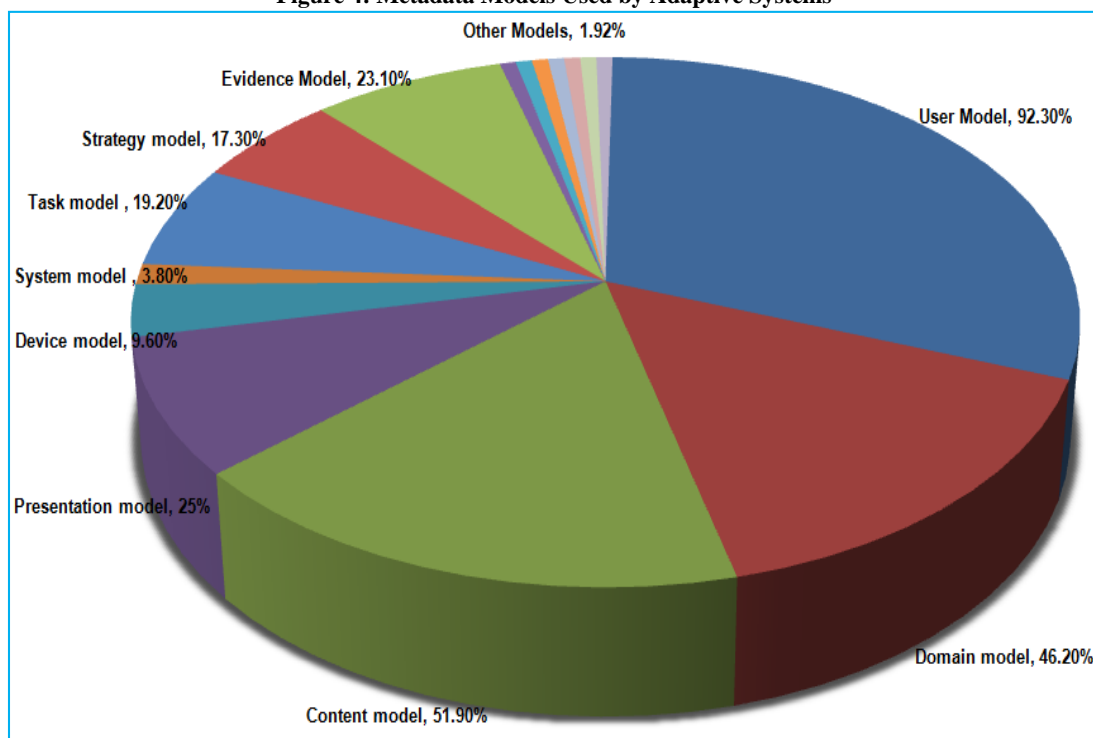


The response to ‘what was improved by adaptivity’ was very good. Only one researcher stated the experiment they conducted showed nothing was improved. We categorized these responses into four groups:

- i) Personalisation because they tailor what is presented to that individual learner:
 - Personalization
 - Personalised books were generated based on student model.
 - Yes the personalised search results were more relevant to users. The number of failed searches was reduced, the time to complete searches reduced etc.
 - Adapted instructional design and personalized guidance to students in terms of recommendations.
- ii) TEL because they lead to better learning through presentation of most relevant learning material:
 - Sequencing of the learning materials navigation support selection of tasks to work at feedback and selection of tasks based on students' current knowledge and preferences.
 - The aim was to provide students with tailored education in the way of choosing an appropriate level of difficulty. Additionally the system provides a course generator for different learning scenarios, such as preparing for an exam. Adaptivity is also included in the way of providing color-flag feedback and the availability of hints, increasingly offering more information about the correct solution or the path towards the solution.
 - Privacy protection.
 - Tailoring Content to specific users.
 - Content that is presented to users.
 - Awareness and learning support.
 - Adaptation of: - background of student’s (course and discipline) - navigation preferences - knowledge level.
 - Recommendations.
 - Presentation and Interaction.
 - The content and the navigational guidance provided to students depending on personal features, actions and current context (device, time and physical location).
 - The recommendations about what to learn next.
 - The learning outcome and improved English learning at early ages (3 to 6 years old).
 - Students solved a science problem-solving scenario more quickly, and received pedagogical supports that were tailored to their curricular knowledge and problem-solving behaviors.
 - Studying behaviour; engagement with lifelong learning; mathematical generalization; theory-aware learning design.
 - The ability of users with a tremor disorder to separate deliberate motions from involuntary motions.
 - Navigational abilities of the robot, ability to escape from traps, speed with which robot could complete the task. Transferability between different robotic platforms

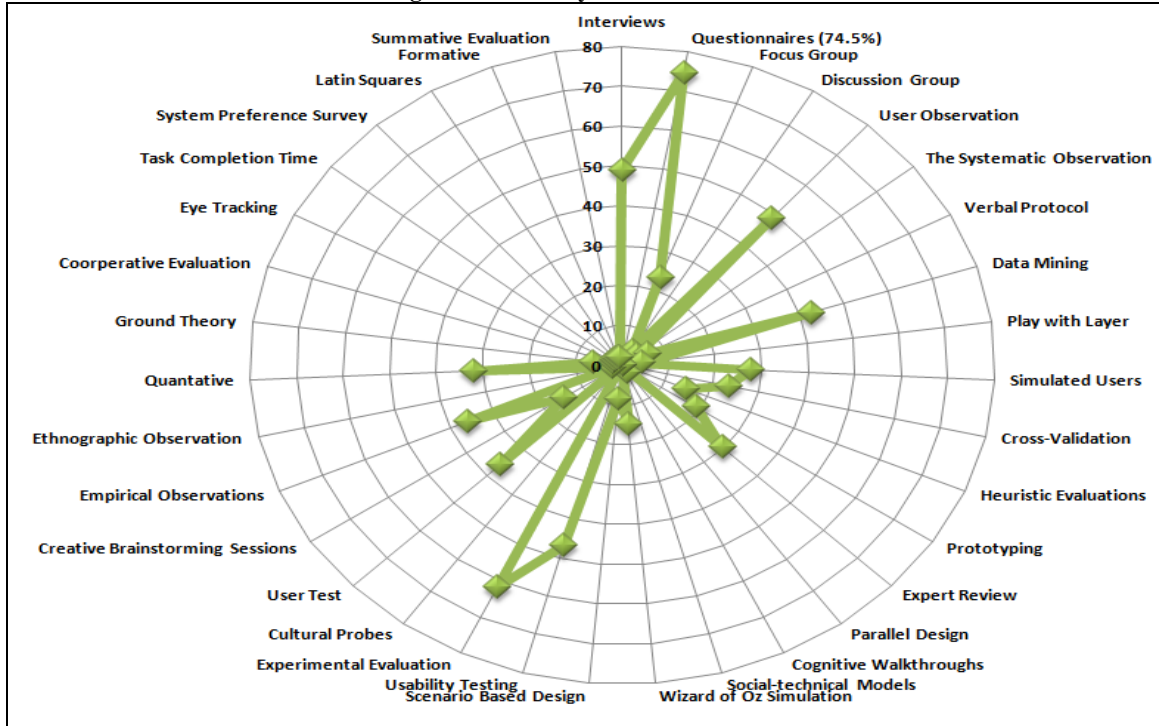
- The sequence of materials shown to students.
 - Students can see the model of their level of understanding in a range of topics, and make informed decisions about their learning. They can also use this information as a basis for peer collaboration.
 - Useful adaptation, in the form of link annotation/hiding and the conditional inclusion of fragments.
 - Some personalised teaching (e.g. matching the information to the learner (e.g. SASY's demonstrators), reducing the amount of information displayed (e.g. Locator) and interpreting information about the user differently (e.g. Locator).
 - Student's intelligent skills.
- iii) User satisfaction:
- The users were more satisfied with the system, but only when it explained the adaptation to the user.
 - User efficiency, effectiveness and satisfaction
- iv) Results output:
- The ranking of search results obtained from search engines was adapted to user needs.
 - The quality of the results and explanatory facility.
 - Search engine results (i.e. the system was particularly focused on query disambiguation).
 - Accuracy of recommendations.
 - Training plan, scenario, setting; intensity and amount of exercise required; recommendations.
- v) Time:
- The length of the test and report on student knowledge.
 - Reduce required time alternatively by keeping the same number of questions.
 - Increase test reliability.

Figure 4: Metadata Models Used by Adaptive Systems



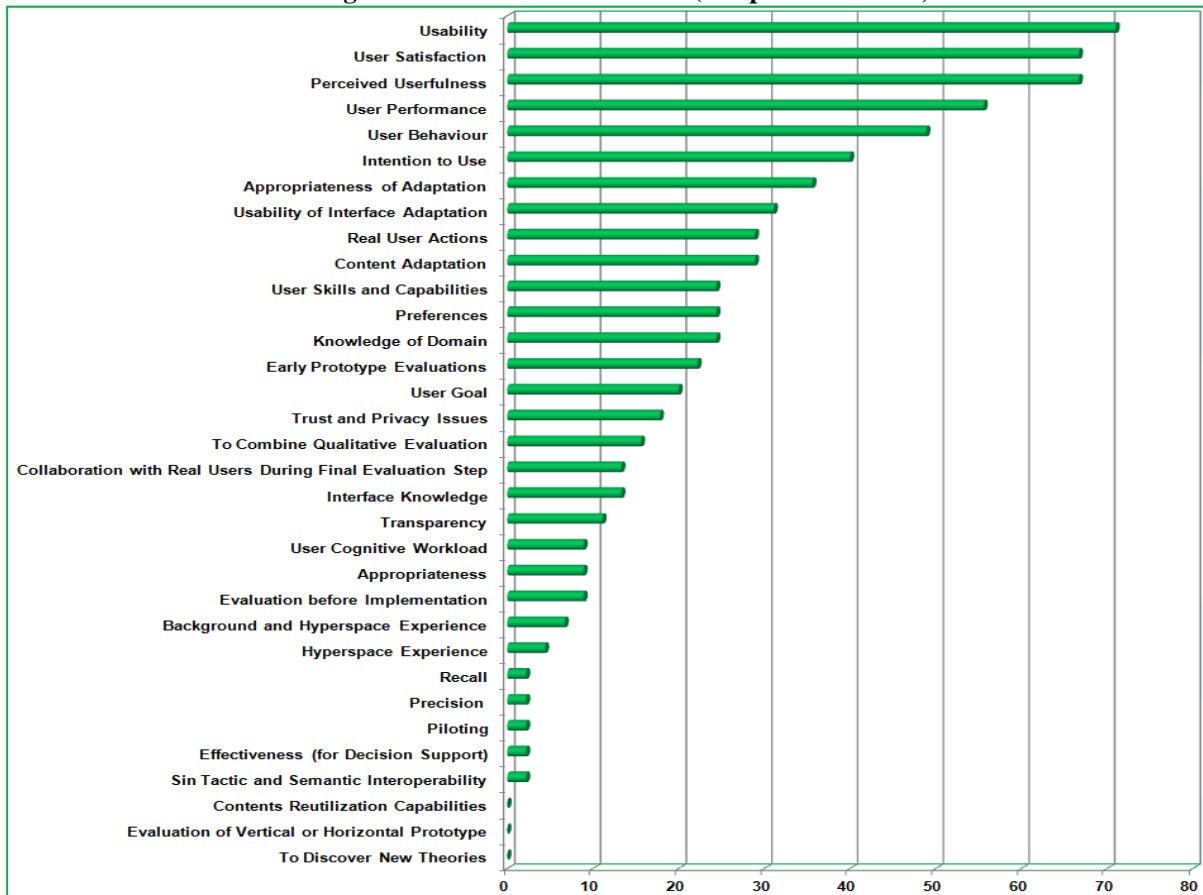
The most commonly used metadata model of adaptive systems was user model, followed by content and domain models (Figure 4). The evaluation of these models was rarely done. 9 people responded to this question and 82 skipped it. This was a clear indication of lack of evaluations of such models. The methodologies for evaluating adaptive TEL systems are generally borrowed from the methodologies used in HCI and by those utilized for the evaluation of the information selection process (Gena, 2005). In the study 60 different types of evaluation methods (Figure 5) for adaptive systems were reported. Most commonly used methods were questionnaires followed by experimental observations, interviews and user observation respectively.

Figure 5: Summary of Evaluation Methods



It is significant to ensure that the correct measurement criteria and metrics are used. A total of 43 measurement criteria also known as adaptive variables were reported (Figure 6). The most commonly used were usability, user satisfaction, perceived usefulness and user performance respectively.

Figure 6: Measurement Criteria (Adaptive Variables)



A total of 32 metrics were reported. Mostly commonly used metrics were accuracy of recommendations, precision, accuracy of retrieval and reliability metrics (Table 2).

Table 2: Commonly Used Metrics

Metrics Name	Response %	Response Count
Accuracy of Recommendations	64.5	20
Precision	61.3	19
Accuracy of Retrieval	35.5	11
Reliability Metrics	19.4	6
Behavioural Complexity	12.9	4
pIA: Performance Influence on Adaptivity	12.9	3

The analysed educational dataset from the survey was used to populate the recommendation service proposed in section 3. A brief summary of potential benefits of personalisation and personalised learning is presented in section 2.2 below

2.2 Overview of Personalisation and Personalised Learning

Personalisation can be based on many different attributes. These include: i) User age, user disability, subject-pre-requisites, user role, user motivation, user language, preferred modality (speech, video), user prior knowledge, user competencies, user experiences/history, user emotion, user objectives, user emotion, user preferences, user interests, user goals and user behaviour (Wade, 2009), ii) The quality approach, the lifecycle approach, iii) display device relationship to other objects, time, performance, level of control, activity, process rules, interaction (with disciplines, group membership, group activity, deadline. One of the core issues in learning is the personalisation of the learning experience. It is widely recognized that effective and efficient learning need to be individualized, personalised and learner controlled. The dimensions of personalisation include: content models, subject domain models, environment models, service models, user and cultural models, activity models, implicit and explicit model triggers, artificial intelligence and non-artificial intelligence approaches. Multi dimensional enables: personalised (collaborative) tasks, personalised situational simulations, personalised games, personalised mobile collaboration, personalised social networking and personalised community support. Personalisation for the end user improves: i) user efficiency, user effectiveness, iii) user satisfaction and access to the otherwise marginalized. It is significant to review aspects of the personalisation that needs to be evaluated (e.g. quality of the user modelling, performance of different adaptation approaches, knowledge gain from using the personalised system or overall end user experience), several evaluation techniques need to be combined and executed differently. The recommendation technique proposed in section 3 presents a novel way of executing this task.

Personalised learning is a form of learning which takes place in a learning environment specifically customized to the individual learner. According to Halm (2006), personalised learning *“meets the needs of the individual learner by providing the best method of learning based on their personal interests, learning styles, motivation and learning objectives”*. Personalised learning activities are similar to learning activities but have: i) added advantage that the content, services and workflow are dynamically adapted to benefit the learners, ii) adaptivity is based on different ‘dimensions’ (Wade, 2009), iii) combination of adaptive selection and sequencing of multimedia content with adaptive selection and sequencing of user centric services. These activities have been widely accepted by several researchers as a means of providing greater learning engagement and enhancement, of the learning activities. Furthermore there are many ways in which personalised learning can be supported: these include: i) assessment for learning (e.g. setting personal targets, effective feedback to the learner effective use of data to plan learning, improved transition and transfer, peer set and self assessment, pedagogy ii) effective teaching and learning (e.g. lessons in learning, mentoring strategies, wider teaching repertoire, interactive, inclusive teaching programmes, ICT across the curriculum, (Pollard and James, 2004), iii) curriculum entitlement and choice (e.g. pupil choice for study and learning, models and materials for catch-up and extension, creating time for tailored curriculum, flexibility leading to relevant qualifications for all, iv) leadership and management focus on teaching and learning, workforce organized appropriately, buildings facilitators of personalised learning, clear behaviour and attendance policies, v) beyond the classroom (e.g. parental involvement, learning in community context, co-ordinated services in/out of the schools to support the whole child, business partnerships, networks and collaboration. Common dimensions of personalised learning are ownership, participation, diversity, reflection (Verpoorten et al., 2009).

In conclusion personalised learning is very significant to each individual learner. Very few educational datasets for evaluating adaptive TEL systems exist; the data collected from this survey is very important to the adaptive hypermedia, recommender and TEL scientific communities. Evaluators of TEL systems face a difficult challenge in knowing “*how to combine the different evaluation methods, criteria and metrics*” during the evaluation process. The proposed recommendation service in section 3 addresses this challenge.

3. Proposed Hybrid Recommendation Service for Recommending Evaluation Techniques

The proposed hybrid (case-based and knowledge-based) recommendation service is built upon an educational evaluation dataset. It is crucial that software developers and evaluators evade well-known pitfalls and that writers of future evaluation reports increase their empirical value, by reporting the used approaches. In this paper evaluations approaches are considered as any technique, method, set of criteria, tool, checklist or any other evaluation/verification instrument and mechanism which has the purpose of evaluating the quality of learning resources. Several approaches (quality approach, lifecycle approach, combined and a layered evaluation approach, combined four-level and six-level approach, user-centered evaluation approach, empirical approach, utility-approach, collaborative filtering, content-based, demographic, the knowledge-based and hybrid) have been used in evaluating adaptive systems (Ehlers et al., 2005)(Drachsler et al., 2010) (Breitner and Hoppe, 2005) (Mulwa et al., 2011c). The following use case scenario demonstrates the process of recommending an evaluation approach to a software developer or evaluator of a adaptive system:

Suppose User X is a novice user and user Y an expert; both want to use our service and do not which evaluation approach to use. In the initial stage of recommendation, we ask both users if they need recommendations for evaluating a new or pre existing system. If the user selects a new system, they are prompted to enter the name of the system otherwise they can select from a list of existing systems developed from 2000 to 2012. To recommend an approach, the following is considered (evaluation purpose, user’s system characteristics, kind of system/variation type, kind of evaluation results needed, kind of question he/she is answering/has to answer and the kind of information input (resources/evaluation) required. Table 3 presents the factors considered during the implementation of the recommendation service for an evaluation approach.

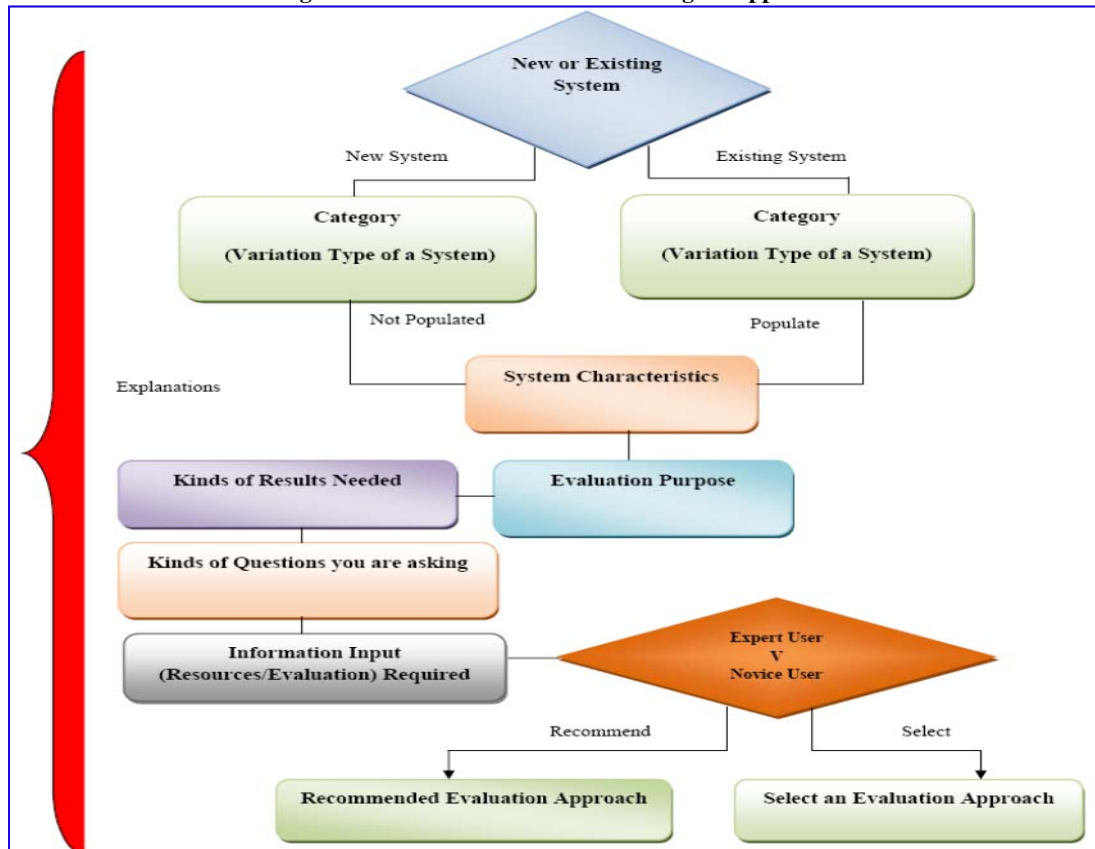
Table 3: Factors considered when recommending an evaluation approach for Adaptive TEL System

Factor	Normalized Value	Weight	Explanations
Number of publications in which the approach was used	NP/TP	1	“Because the approach has been used NP times out of TP times in the literature (2000-2012)”
The types of (venues) publications that the approach has been used in (e.g. journal, conference and workshop	N_V/T_V	1	“Because the approach appeared in (N _j journals), and (N _c conferences), and (N _w workshops)”
How many adaptive systems belonging to the same variation type(category) have been evaluated using the approach	N_{SV}/T_{SV}	2	“because, out of the T _{sv} systems which belonged to the “V” variation Type, N _{sv} of them have been evaluated using the approach”
Give an extra weight to the approach according to its association with the selected evaluation purpose	N_{EPV}/T_{EPV}	2	“Because, out of your T _{EPV} selected evaluation purposes, N _{EPV} of them are associated with the approach”

P: Publications (evaluation study) S: System V: Variation Type (Categories) EPV: Evaluation Purpose N: Number of... T: Total number of... N _V : Venue score of the evaluation approach T _V : Total venue score of all publication	NP: Number of publications published that used the evaluation approach TP: Total number of publications in the database repository N _{SV} : Number of systems that belong to a given variation type (v) that were evaluated using the same evaluation approach T _{SV} : Total number of systems that belong to a given Variation Type (v) N _{EPV} : Number of selected evaluation purposes that are associated with the evaluation approach T _{EPV} : Total number of evaluation purposes selected by the user (i.e how many check boxes the user checks on the screen)
--	--

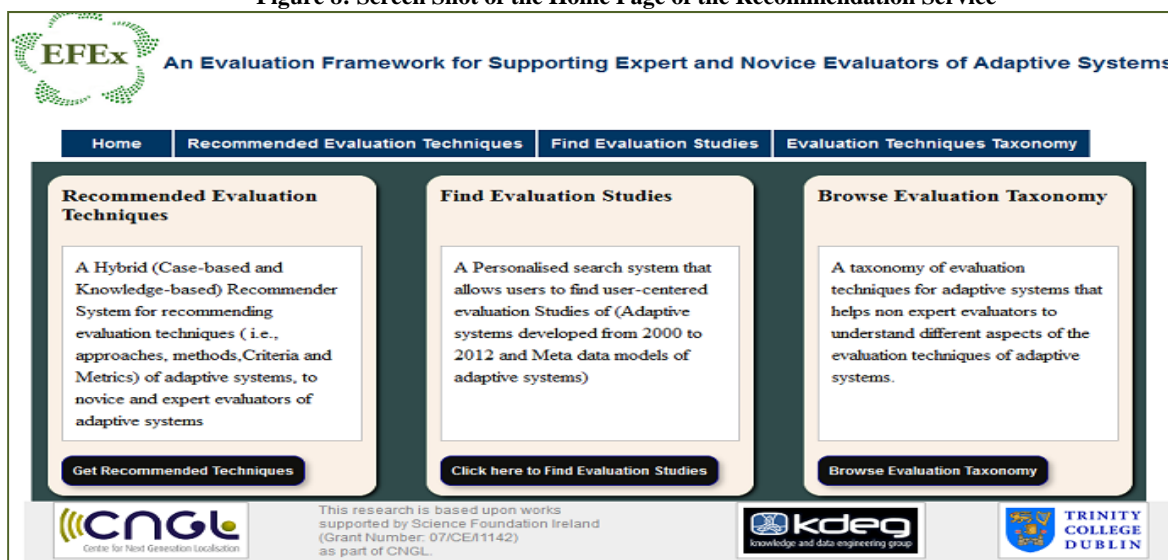
The process of recommending an evaluation approach is depicted in Figure 7. Throughout this process both users are provided with explanations as to why each factor was taken into consideration and why that approach was recommended. These explanations are important because they provide: i) transparency, validity, trustworthiness, ii) persuasiveness, effectiveness, efficiency, iii) satisfaction, relevance, iv) comprehensibility and education (Tintarev 2007, Masthoff 2007).

Figure 7: The Process of Recommending an Approach



Users are provided with a web-based user interface (Figure 8) of the recommendation service. Information is translated into user's choice of language.

Figure 8: Screen Shot of the Home Page of the Recommendation Service



4. Conclusion

In this paper we have presented a study of evaluations of adaptive TEL systems. We have also raised current concerns, issues and challenges encountered by evaluators of these systems. It is not only important to evaluate the TEL system but also to ensure correct evaluation approach and techniques are used and finally value the integration within the context where the learning will take place. Recommending the most appropriate evaluation approach is very significant. Currently we are conducting user trials of the recommendation service using task-based and interview evaluation method.

Acknowledgement

This research is based upon works supported by Science Foundation Ireland (Grant Number: 07/CE/I1142) as part of research being conducted at the Centre for Next Generation Localization (www.cngl.ie). The authors are grateful for the suggestions of the reviewers of this paper.

References

- Breitner, M. & Hoppe, G. 2005. A glimpse at business models and evaluation approaches for e-learning. *E-Learning*, 179-193.
- Drachler, H., Bogers, T., Vuorikari, R., Verbert, K., Duval, E., Manouselis, N., Beham, G., Lindstaedt, S., Stern, H. & Friedrich, M. 2010. Issues and considerations regarding sharable data sets for recommender systems in technology enhanced learning. *Procedia Computer Science*, 1, 2849-2858.
- Ehlers, U. D., Goertz, L., Hildebrandt, B. & Pawlowski, J. M. 2005. *Quality in e-learning: use and dissemination of quality approaches in European e-learning: a study by the European Quality Observatory*, Office for Official Publications of the European Communities.
- Gena, C. 2005. Methods and techniques for the evaluation of user-adaptive systems, *The Knowledge Engineering Review*. 20, 1-37.
- Jameson, A. 2009. Adaptive interfaces and agents. *Human-Computer Interaction: Design Issues, Solutions, and Applications*, 105.
- Lavie, T., Meyer, J., Beugler, K. & Coughlin, J. 2005. The evaluation of in-vehicle adaptive systems, *User Modeling: Work on the EAS*. 9-18.
- Lawless, S., Mulwa, C. & O'Connor, A. 2010. A Proposal for the Evaluation of Adaptive Personalised Information Retrieval. *In: Proceedings of The 2nd International Workshop on Contextual Information Access, Seeking and Retrieval Evaluation*, 28th March 2010 Milton Keynes, UK. CEUR-WS.org 4.
- Markham, S., Ceddia, J., Sheard, J., Burvill, C., Weir, J., Field, B., Sterling, L. & Stern, L. 2003. Applying agent technology to evaluation tasks in e-learning environments, pp 16–17.
- Missier del, F. & Ricci, F. 2003. Understanding recommender systems: Experimental evaluation challenges. 31-40.
- Mulwa, C., Lawless, S., Sharp, M. & Wade, V. 2011a. The evaluation of adaptive and personalised information retrieval systems: a review. *International Journal of Knowledge and Web Intelligence*, 2, 138-156.
- Mulwa, C., Lawless, S., Sharp, M. & Wade, V. 2011. A Web-Based Framework for User-Centred Evaluation of End-User Experience in Adaptive and Personalized E-Learning Systems, *IEEE Computer Society*, 351-356.
- Mulwa C., Lawless., S., Sharp., M., Wade., V. & Arnedillo-Sanchez., I. 2010. Adaptive Educational Hypermedia Systems in Technology Enhanced Learning: A Literature Review. *In: SHERIDAN (ed.) Association for Computing Machinery's Special Interest Group for Information Technology Education*. Central Michigan University at Midland MI USA ACM Proceedings.
- Mulwa, S. L. C., Sharp, M. & Wade, V. 2011c. The Evaluation of Adaptive and Personalised Information Retrieval Systems: A Review. *International Journal of Knowledge and Web Intelligence (IJKWI)*, 20.
- Pollard, A. & James, M. 2004. *Personalised learning: A commentary by the teaching and learning research programme*.
- Tintarev, N. & Masthoff, J. 2009. Evaluating Recommender Explanations: Problems Experienced and Lessons Learned for the Evaluation of Adaptive Systems. *User Modeling, Adaptation and Personalization*. Trento, Italy.
- Verpoorten, D., Glahn, C., Kravcik, M., Ternier, S. & Specht, M. 2009. Personalisation of learning in virtual learning environments. *Learning in the Synergy of Multiple Disciplines*, 52-66.
- Wade, V. 2009. Challenges for the Multi-Dimensional Personalised Web. *UMAP Trento*.
- Weibelzahl, S. & Weber, G. 2002. Advantages, opportunities and limits of empirical evaluations: Evaluating adaptive systems. *KI*, 16, 17-20.