EMPI: A questionnaire based method for the evaluation of multimedia interactive pedagogical software

Stéphane Crozat, Philippe Trigano, Olivier Hû

EMPI: A questionnaire based method for the evaluation of multimedia interactive pedagogical software.

Stéphane CROZAT, Olivier HU, Philippe TRIGANO
UMR CNRS 6599 HEUDIASYC
Université de Technologie de Compiègne - BP 20529
60206 COMPIEGNE Cedex - FRANCE
Email: Stephane.Crozat@utc.fr, Olivier.Hu@utc.fr, Philippe.Trigano@utc.fr

***Abstract**: We submit a method to help in evaluating the multimedia learning software. We aim to assist the users (mainly teachers and students) to make a choice between the large range of software actually available. Our process is to divide the software analysis into six main themes: the general feeling, the technical quality, the usability, the multimedia documents, the scenario and the didactics. Each of these themes is sub-divided into criteria, sub-criteria and questions. The whole forms a hierarchical questionnaire that allows marking software through various aspects, in order to compare it to other software or with a determined pedagogical context. This paper presents the detailed structure of the questionnaire, through the criteria which compose it, along with some examples of questions, and, to end with, some aspects of the software we are making in order to bring the method into operation.*

***Keywords**: Multimedia, Software Evaluation, Instructional Context, Ergonomics.*

1 Introduction

We can detect an increasing infatuation in institutions and families with the use of new technologies and multimedia in an educational context. They have to be integrated into schools, into houses, they have to be used by children and by adults. Listening to these voices we should use new didactical technologies in all situations for all people. However, watching what actually happens, we have to state that new technologies are often ignored, forgotten, subused and indeed rejected. We do not think that the technology is to be rejected in itself, there is no reason why it should not find a place close to the book, the traditional teaching and the firms training (let us note that there are no reasons why it should either replace them!). However we think that the relative failure of multimedia learning software is due to their poor quality, compared to what they could offer and what the public expects them to offer. Design mistakes, poor contents, unusable interfaces, bad use of multimedia potential are samples of usual failings. Nevertheless these alternative and complementary ways of teaching are particularly advantageous in specific cases, such as distance learning, along-the-life learning, very heterogeneous skills in classes, children helping,…

On one hand, one of the problems linked to that observation is the difficulty of choice of a product, and more widely the problem of evaluation: How to know if such software is better than another regarding the contents? How to estimate if the interface would be easy to use? How to find the most adapted software for a requested situation? Does the learning software really use the potential of multimedia technology? To answer these questions, we need tools to characterise and evaluate the multimedia learning software, against relevant criteria. The one we submit is a helping method for the Evaluation of Multimedia, Pedagogical and Interactive software (EMPI).

We shall first present the method and the linked questionnaire, then we will develop the six main themes, and in the last part we shall briefly present a software package used to implement the method and the validations we made on it.

2 Method principles

2.1 Position

Multimedia learning software evaluation comes from two older preoccupations: The evaluation of pedagogical aids (scholar's books for instance) [1] and the software and human-machine interfaces (mainly in industrial context) [2]. We shall try to adapt both into the more specific field of learning software. The tool we propose is expected to be general, however we had to restrict this wide field in some aspects. The evaluation should be done by the user, the de-

cider of the pedagogical strategy, or the manager of a learning centre. We also want to deal directly with the software (in terms of usability, multimedia choices, or didactical strategy) not with its impact on users. Our method is expected to be used on manufactured products, not in a fabrication process. Nevertheless we shall discuss this last point in our conclusion.

2.2 Questionnaire structure

We oriented our researches towards several areas: computer sciences, ergonomics and multimedia first, but also other areas linked to cognitive sciences, social sciences, artistic sciences,… Faced with the complexity of such ambitions, we adopted an iterative approach: Firstly, we began with usability oriented studies, we then worked on didactics, and ended with multimedia aspects. Each time our method was to extract criteria from the related literature, to test these criteria, to integrate them into a prototype and to evaluate them in real situation. After each evaluation we could begin a new cycle, integrating new aspects we thought relevant. At each step the initial method and the previous criteria were also changed, in order that the new studies introduced new constraints and ideas.

Today we hope to have reached a stable structure. Nevertheless, there is no doubt that further evaluations will continue to modify the questionnaire, but only in more specific aspects. So, finally, we decided to divide the global evaluation into six main approaches, or themes:

- The **general feeling** takes into account what image the software offers to the users.
- The **computer science quality** allows the evaluation of the technical realisation of the software.
- The **usability** corresponds to the ergonomics of the interface.
- The **multimedia documents** (text, sound, image) permit the evaluation of the contents presentation.
- The **scenario** deals with the writing techniques used in order to design information.
- The **didactical** module finally inspects the pedagogical strategy, the tutoring, the learningsituation,…

Each of these themes is sub-divided in criteria, sub-criteria and questions. This hierarchical structure allows variable depth inspection, depending on the skills and the wills of the evaluator.

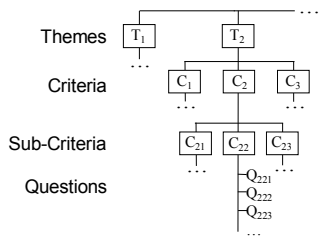


Figure 1: Questionnaire structure

An evaluator very competent in the ergonomics domain would not need to deepen the criteria of this theme, but he could expect to be strongly driven for the didactical aspects. In the same way, one could not be interested in deepening the criteria of personalization (see p.3) if the software would be used in very punctual contexts, without time for adaptation.

2.3 Questionnaire characteristics

Our method is founded on a questionnaire that allows the marking of each criterion, at each level. That means that the evaluator can directly evaluate each criterion, instinctively, or go deeper accessing corresponding sub-criteria, then questions. The evaluating system manage two kind of marks: the **instinctive marks** (++/+/=/–/– –) that are directly attributed to the criteria by the evaluator, and the **calculated marks** that are attributed to the criteria by the software using the answers the evaluator gave to the questions. A confrontation is possible between the marks, using the consistency rating (that determines if the instinctive marks are coherent between themselves) and the correlation rating (that indicates if the instinctive and calculated marks converge).

For the calculated marks we use an **exponential marking** in order to have the defaults underlined:

Example: *Did you ever happen not to know what to do to keep on using the software?Always (-10) / Often (-6) / Sometimes (0) / Never (+10)*

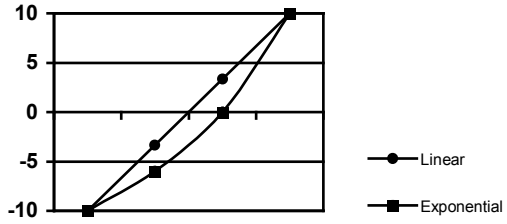


Figure 2: Exponential marking

Some questions are subdivided in two phases: A first one to **characterise** the software's situation, and a second one to evaluate the relevance of this situation. For instance, in order to evaluate the structure of the software, we will first deter-

mine what kind of structure is concerned (linear, arborescent,…) and then if it is a correct one.

The evaluator, with a synthesis of the instinctive and calculated marks and the correspondent ratings, is given a **final mark** by the evaluating system. But the human evaluator keeps ultimately the capacity of judging the final mark of each criterion.

A **structured and contextual help** is provided for each criterion and question, in order to have the most objective evaluation. This help allows questions reformulation, concepts definition, theoretic fundaments explanation and some characteristic examples.

The **weight of questions** on a criterion can be either essential or secondary, to express the fact that some aspects or defaults are more important than others.

3 Themes description

In this part we shall develop each theme. The whole criteria list (Figure 3) and some examples of questions (Figure 4) are proposed in annexes.

3.1 General feeling

Several experiences we had drove us to the idea that software, especially multimedia software, provides a general feeling to the users. This feeling comes mainly from graphical choices, music, typographic, scenario structure,… The important fact is that the utilisation of the software is greatly influenced by these feelings. For instance we could think that the software seems complex, or attractive, or serious,… And the impressions the user gets deeply affect the way he learns. We studied various fields, such as visual perception theories [3], image semantic [4], musicology [5], cinematography strategies [6],… With these theories and the practical experiences we had, we managed to submit a list of criteria. We shall specify that this theme is particular in the following senses: the criteria are provided by opposite pairs; they are expected to be neutrals, in order to describe the feelings, not to judge them directly; there is no sub-criteria level, nor questions directly linked to the criteria. In fact, we want the evaluator to characterised the general feeling by using the submitted criteria, in order to determine if it is adapted or not to the pedagogical context.

3.2 Technical quality

This part of the questionnaire concerns the classical aspects of software engineering. It was not our main concern to deeply research on this subject, since previous researches have already investigated these areas.

3.3 Usability

Usability evaluation has been widely studied, especially within the industrial context [7,8,9,10]. The ones we chose are mainly based on INRIA criteria [11]. They are more deeply described in [12,13].

3.4 Multimedia documents

Texts, images and sounds are the constituents of the learning software. They are the information vectors, and have to be evaluated for the information they carry. But the way they are presented is an important point, because it will influence the way they are read. In this part we also inspect the relevancy of the choices made in terms of redundancy and complementarily of media. To build this part of the questionnaire, we had to explore [14] various domains, such as the pictures' semantics [15], the textual theories [16], the didactical images works [17], the photography [18], the audio-visual [19],…

3.5 Scenario

We define the scenario such as the particular process of designing multimedia documents in order to prepare the act of reading. The scenario does not deal directly with information, but with the way they are structured. This supposes an original way of writing, dealing with non-linear structure, dynamic data, multimedia documents,… Our studies [14] are oriented toward the various classification of navigation structures [20,21], and the fiction integration in learning software [22].

3.6 Didactics

Literature offers plenty of criteria and recommendations for the pedagogical application of computer technology, for instance [23,24,10,25]. We also used more specific studies, such as reflections on interaction process [26], or practical experiences [27]. This last part of the questionnaire is expected to evaluate the specific didactical strategy of the software. Our goal is not impose one or another strategy, saying it is the better one. This normalising approach can not be applied (whereas it was possible for ergonomics or technique), for two main reasons: We do not have enough experience with learning software to impose a way of doing things and the evaluation of a didactical strategy is totally context dependent. That means that our method is not able to directly evaluate the criteria, but what it can do is giving the evaluator a main grid to determine on each point what kind of strategy is

chosen and if this is relevant regarding the particular context of the learning situation.

4 Validation

The use of the questionnaire we describe, and moreover the exploitation of results need to be implemented in a software version in order to be really effective. Such software is actually being made. We already ended a first prototype, realised as a database with Access.

Several versions of the questionnaire have been successively set up. The first researches, centred on ergonomics, revealed the necessity to take into account didactics and multimedia aspects. Various validations have been made, mainly on the ergonomic module. For the first one, ten evaluators used thirty learning software. It enabled improvement of the usability module. We also began to consider the necessity of other evaluation themes. The second validation permits comparison of forty-five evaluations of the same software, using a stability rating. Here could be underlined some weak parts of the questionnaire. The third study was mainly centred on the comparison between our method EMPI and the MEDA method, the only commercial evaluating method based on questionnaire. We shall refer to other articles for the details of these studies, [13] for instance.

We have a new validation program in order to extend the experiments to all of the themes of the formerly described questionnaire. Particularly, we plan to make another large experiment with fifty evaluators, and to distribute the prototypes for validation on site. However we said in our introduction that our method aimed at both users and prescriptors, we have to point out that our validations experiments only concerned users until today. For practical reasons of availability we used students for the preliminary tests. In the further experiments that we plan to carry out, we will include teachers in the prescriptors' role.

5 Conclusion and perspectives

5.1 Evaluations, evaluators, contexts

Submitting a generalist tool that allows the global evaluation of any software used in educational context is something ambitious. Some would say impossible, some would say indispensable, and probably both would be right. In front of difficulties such as the evaluation of subjective aspects (opportunities of colour choices for instance) or evaluation of contextual aspects (contents of the software) we had to adopt an humility attitude. The method we submit is not really able to evaluate, and can only help the evaluator by its systematic approach. But toward each of this specific criterion the only one who can judge is the human being. This implies various restrictions we have to deal with. For instance any evaluator will not be able to do the whole evaluation, but only evaluators who know the pedagogical context, able to have enough distance to globalise their vision. Our modularly approach is a way of adapting each evaluation to its evaluator and its domain. The present version of the method does not offer any support to specify the relevant criteria for a determined context. The deeper research and experiment we make will help us in determining the precise limits of our method. Particularly, we want to determine the skills an evaluator should gather for each theme (hoping that usability and technical quality themes, at least, will not need much). It could be greatly helpful in the interpretation of evaluation results. We also try to select which criterion is adapted to which context (hoping to stay as global as possible).

5.2 A standardisation tool

If a part of the education community adopts the method, our criteria could be used as a standard reference. In front of new software, one would have to apply the questionnaire in order to determine their weakness and strength, and then compare with what he is looking for in his particular context. But it could be very helpful for the rest of the community to share these results. It would allows comparison between different evaluations, detecting best software, discussion on particular aspects, saving time for widely evaluated software,… Let us hope it could also help designers in taking into account some criteria they do not care very much at the present time.

5.3 From evaluation to conception

The direction of our research is influenced by the result that using the method helps understanding and grasping the concepts we use to evaluate learning software. That means that an experimented evaluator should no longer need the method, for he acquired the knowledge organised in it. He should just use the global criteria grid, in order to be sure not to forget any point. However, his eye would be trained enough to directly know if such software is good or not against each criteria. Of course this intuition we have needs to be proved and measured, but we already plan to orient the method in order to promote this tendency. For instance we shall

try to explicit to the user the knowledge we use, rather than giving him ratings and markings without explanations. To sum up we could say that EMPI method helps *reading* learning software. But we also remark that it helps *writing* ones (for writing can not be separated from reading…). Concretely, we begin a new research activity, complementary, which purpose is to reverse the evaluation criteria, in order to submit design recommendations and methods for conception.

References

[1]Richaudeau F., Conception et production des manuels scolaires, Paris, Retz, 290p, 1980.

[2]Kolski C., Interfaces Homme-machine : application aux systèmes industriels complexes, Hermes, 1997.

[3]James J. Gibson, «The ecological approach to visual perception», LEA, London, 1979.

[4]Claude Cossette , «Les images démaquillées», Riguil Internationales, 2ème édition, Québec, 1983.

[5]Michel Chion, «Musiques: Médias et technologies», Flammarion, 1994.

[6]Francis Vanoye, Anne Goliot-Lété, «Précis d'analyse filmique», Nathan, 1992.

[7]Ravden S.J., Johnson G.I., Evaluating usability of Human-Computer Interfaces : a practical method. Ellis Horwood, Chichester, 1989.

[8]Vanderdonck J., Guide ergonomique de la présentation des applications hautement interactives, Presses Universitaires Namur, 1994.

[9]Senach B., Evaluation ergonomique des interfaces Homme/Machine : une revue de la littérature. Rapport INRIA, Sophia-Antipolis, n°1180, Rocquencourt, mars 1990.

[10]MEDA, «Evaluer les logiciels de formation», Les Editions d'Organisation, 1990.

[11]Bastien C., Scapin D., Evaluating a user interface with ergonomic criteria. Rapport INRIA n°2326 Rocquencourt, 1994.

[12]Olivier Hû, Philippe Trigano, «Proposition de critères d'aide à l'évaluation de l'interface homme-machine des logiciels multimédia pédagogiques», IHM'98, Nantes, septembre 1998.

[13]Olivier Hû, Philippe Trigano, Stéphane Crozat, «E.M.P.I.: une méthode pour l'Evaluation du Multimédia Pédagogique Intéractif», NTICF'98, INSA Rouen, novembre 1998.

[14]Stéphane Crozat, «Méthode d'évaluation de la composition multimédia des didacticiels», Mémoire de DEA, UTC, 1998.

[15]Yveline Baticle, «Clés et codes de l'image: L'image numérisée, la vidéo, le cinéma», Magnard, Paris, 1985.

[16]Jack Goody, «La raison graphique: La domestication de la pensée sauvage», Les Editions de Minuit, 1979.

[17]Joan Costa, Abraham Moles, «La imagen didáctica»,Ceac, Barcelone, 1991.

[18]Henri Alekan, «Des lumières et des ombres», Le sycomore, 1984.

[19]Pierre Sorlin, «Esthétiques de l'audiovisuel», Nathan, 1992.

[20]Alain Durand, Jean-Marc Laubin, Sylvie Leleu-Merviel, «Vers une classification des procédés d'interactivité par niveaux corrélés aux données», H²PTM'97, Hermes, 1997.

[21]Patrick Pognant, , Claire Scholl «Les cd-romculturels», Hermès, 1996.

[22]Patrick Pajon, Olivier Polloni, «Conception multimédia», cd-rom, CINTE, 1997.

[23]Dessus P., Marquet P., Outils d'évaluation de logiciels éducatifs. Université de Grenoble. Bulletin de l'EPI. 1991.

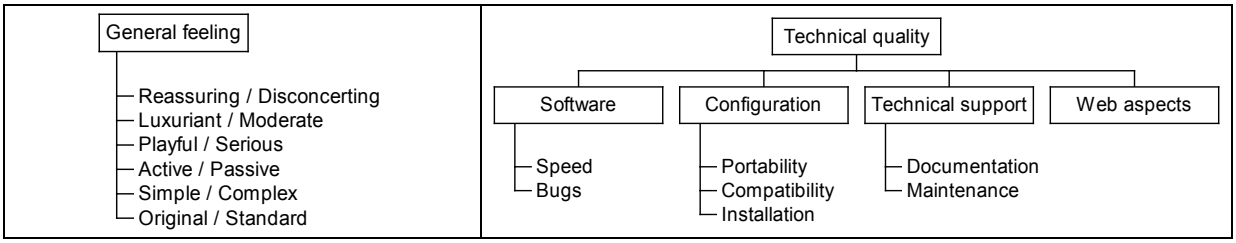
[24]P. Marton, «La conception pédagogique de systèmes d'apprentissage multimédia interactif: fondements, méthodologie et problématique», Educatechnologie, vol.1, n°3, septembre 1994,

[25]Park I., Hannafin M.-J., Empiically-based guidelines for the design of interactive media, Educational Technology Research en Development, vol.41, n°3, 1993

[26]Martial Vivet, «Evaluating educational technologies: Evaluation of teaching material versus evaluation of learning?», CALISCE'96, San Sebastian, juillet 1996.

[27]Hélène Perrin, Richard Bonnaire, «Un logiciel pour la visualisation de mécanismes de gestion des processus du système UNIX», NTICF'98, INSA Rouen, novembre 1998.

Annexes



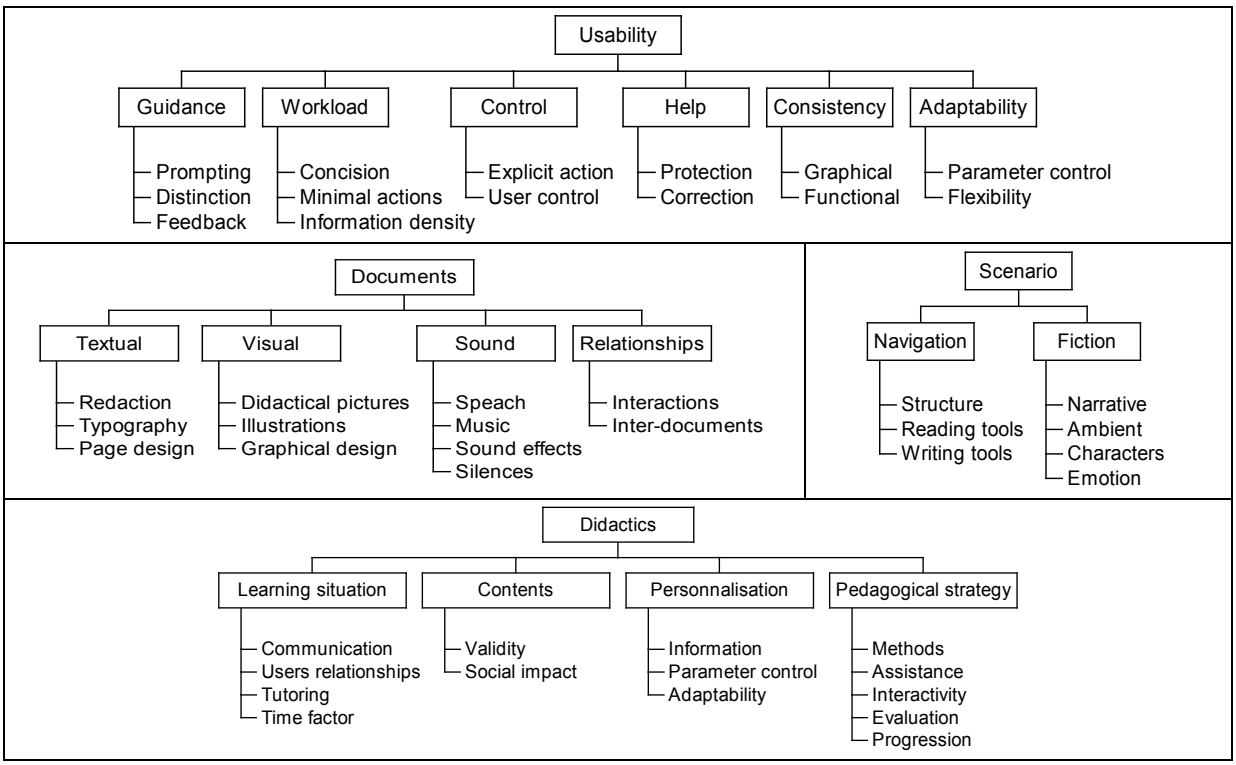


Figure 3 : Criteria

Technical quality / Software / Bugs	
Question Did the software ever produce fatal errors while using?	**Answers** Often / Sometimes / Once / Never
Indications Mistakes in the software design, or incompatibilities between the software and some operating systems can lead to technical errors. Such errors, or bugs, generated by the system have to be distinguished from users' errors foreseen by the software. Examples of bugs are the impossibility of using a command (whereas it should be possible), the loss of mouse or keyboard control, sudden changes in the screen display,… Fatal errors are errors that induce the software stop, or even worth the operating system. In these cases there are no possibilities of control by the user, except reloading the software, or the computer! Of course such bugs should never be met in software.	
Usability / Guidance / Feedback	
Question Are the user's actions followed by a system feedback?	**Answers** Always / Often / Sometimes / Never
Indications User's actions can be a mouse click, a selection, a keyboard validation, a data capture,… A feedback can be either visual (button effects, colour changing, cursor changing,…) or sound (beep, various sound effects,…).	

Documents / Relationship / Inter-documents	
Question Characterise simultaneous presentation of visual and textual documents in the learning software?	**Answers** Symbiosis / Redundancy / Complementarity / Indifference / Divergence

Indications

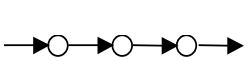
Multimedia software's particularity is to allow several different kinds of media to appear in the same time. The user of such software tries to bind together these different sources of information. The global meaning the user gets is something different from the isolated meaning of each media. Each kind of media and each combination of media imply different way of interpretation.

We propose to distinguish two fundamental relationships: Redundancy when media provide the same information and complementarity when they provide the same one, but in different way. If media are both redundant and complementary we call it symbiosis. Media can also have no interpretable relationship, we call it indifference, or worst can provide contradictory information, we call it divergence.

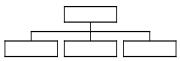
Scenario / Navigation / Structure	
Question What kind of structure is mainly used in the software?	**Answers** Linear / Tree-form / Net-form

Indications

A linear structure is sequential and the user can only control the information flux. This is the usual case for a book or a tape. This structure does not profit from the advantages of the numeric support, but is easier to grasp. A tree-form structure is hierarchic and typically based on menus and sub-menus. It is a compromise between linear and net-form structure. The net-form structure is particular to the numeric support (such as Internet structures). In one hand it allows richer and more adapted readings, because each user is able to have specific path in the net. But in the other hand, it may lose the user, if he is not prepared and guided enough.



Linear structure



Tree-form structure

Net-form structure

Didactics / Pedagogical strategy / Interactivity	
Question Characterise the interactivity level that the system allows the user to reach:	**Answers** Creating / Experimenting / Manipulating / Exploring

Indications

New technologies allow users to act within the software in itself. But we distinguish four levels for this interactivity: Exploring is the lower level, basically present in all learning software permitting navigating, listening, watching, reading, choosing,… Manipulating means users can move, orient, enlarge, detail, combine objects, in order to see them better. For instance this is the case in three-dimensional manipulation of mechanical elements, or manipulation of body parts. Experimentation means participating in interactive simulations, associating data or objects to observe their effects, realising sequenced actions to acquire know-how,… For instance it could be the realisation of physics experiment in virtual laboratory. Creation is the higher level, grouping all the real practices linked to the learning process. It could be the managed used of graphic editor, text processing, music players,…

Figure 4 : Example questions