ReCALL **16** (2): 502–519. © 2004 Cambridge University Press
DOI: 10.1017/S0958344004001727 Printed in the United Kingdom

Enhancing online CALL design: The case for evaluation

DOMINIQUE HÉMARD
London Metropolitan University, Old Castle Street, London E1 7NT, UK
(*e-mail: d.hemard@londonmet.ac.uk*)

Abstract

Whilst the potential of online Computer Assisted Language Learning (CALL) in terms of access, immediacy and exposure is widely acknowledged, as yet remarkably little is known about its impact on the user population, be they language teachers or learners. Indeed, beyond its technological development, online CALL design is still affected by a dearth of reliable evaluative data, critically needed for improving the interface design, implementation and integration. On this premise, this paper proposes to focus on a range of appropriate evaluation methods applied to a recently authored web-based interactive CALL interface, comprising multimedia interactive exercises and self testing, as part of an ongoing research combining Human Computer Interaction (HCI), Java programming and language teaching expertise. This current evaluation adopts a novel approach insofar as its main objectives are not only to assess the usability and usefulness of the interface as well as the student experience, but also to compare the different methods used on the basis of their input and evaluative output. Ultimately, this paper will strive to bring theory and practice closer together by shedding further light on the crucial role of evaluation, the appropriateness of the data it generates and its applicability to online CALL development.

1 Introduction

Nowadays, online CALL developments in Higher Education institutions are more likely to be influenced by prerogatives such as policy, cost and expediency and generated by general purpose platforms than by design principles and learning considerations. For instance, Web-based course authoring templates known as Virtual Learning Environments (VLE) or Learning Management Systems (LMS), which are increasingly, often forcibly, provided to the teaching profession at large, are, in this respect, best exponents of this approach. Firstly because this technology is increasingly designed and 'packaged' for mainstream use by an increasing number of agnostic language teachers and conveniently perceived as a digital converter of language learning material. Secondly, web delivery is largely conceived for "the student as a recipient rather than as a seeker of knowledge and skills" Godwin-Jones (2003:54) and, as such, is considered the ideal vehicle for delivering existing textual material.

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

Even with increased Information and Communication Technology (ICT) expertise and enthusiasm, evidence from designed products and the CALL literature often shows that the interactive transfer of language material on the web is still far too pragmatic and hardly critically evaluated. As a result, the explosion of web technology with its many authoring platforms facilitating courseware development is not producing enough new information about its interactive potential and meaningful use by learners. In other words, and not without some irony, the widespread institutional development of web-delivered language learning material in HE, far from enhancing CALL as a recognised research discipline, is in the process of marginalising it further.

On such a basis, this paper argues that the way forward involves adopting a more reflective and iterative approach towards existing online CALL design and practice supported by the systematic evaluation of the usability and effectiveness of its delivery. In order to do so, a number of relevant user-centred evaluation methods will be applied and compared in a case study intended to assess the necessity for proper evaluative input and output in CALL.

Thus the objectives of this paper are to highlight the role of evaluation by introducing, comparing and contrasting applicable evaluation methods, emphasise its relevance and importance in online CALL design and development and help generate a stronger research-based domain by promoting the use and sharing of evaluative data.

2 Evaluation

2.1 Towards an evaluation framework

Although evaluation is well covered in the HCI literature (see for instance Dix *et al*., 1998; Dumas & Redish, 1999; Karat, 1997; Preece *et al*., 1994, 2002; Shneiderman, 1998) its main parameters are introduced here for clarification.

2.1.1 Evaluating usability

Karat (1997:691) defines evaluation as "the result of a process with a purpose in a context focused on an object". Concretely, this refers to methods, which are employed for the purpose of assessing how a system behaves and is used to carry out certain tasks within a given environment. In so doing, particular emphasis must be placed on the usability of an application or how users can "achieve specified goals with effectiveness, efficiency and satisfaction" (691) by evaluating the designed interaction within a context of use. Thus, such a user-centred approach presupposes that the whole environment within which the software operates or "total system" (693) should be considered including important elements such as user characteristics, the interactive learning context (Squires, 1999: 463) and user support. This wider contextualisation and definition of evaluation is especially pertinent to online CALL practitioners as they are more likely to be involved in designing interactive learning environments using existing authoring systems than developing new software *per se*.

2.1.2 Types and goals of evaluation

There are, roughly, two types of evaluation depending on whether the design or implementation of a system is to be tested. The first approach, also referred to as formative evaluation, is normally conducted by the design team early on but also throughout the

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

whole design process and is intended to ensure that the system behaves as expected through an assessment of its functionality (Dix *et al*. 1998:406). As such, this iterative evaluation relies more on the designer and HCI expertise than on the early involvement of end-users. The second type is summative and focus on the usability of the system as well as the user's experience in order to ascertain if user requirements have been met and to ensure that the impact the design has on the user population is assessed. This form of evaluation, which takes place in the final stages of the design process, relies heavily on user feedback. Thus, evaluation, be it formative or summative, is a crucially important activity of a principled approach to design as it plays a necessary part in assessing initial hypotheses and assumptions (Levy, 1999), checking a design solution against identified problems as well as facilitating and supporting the process of understanding users' needs and satisfaction (Hémard, 2003).

2.1.3 Evaluative data

There are two main categories of measurements linked to and generated by the different evaluation methods according to their varying degree of objectivity or subjectivity. The more objective techniques tend to adopt a numeric approach and produce quantitative data generally analysed using statistical techniques such as values, tables and graphs. The more subjective techniques yield non-numeric, qualitative data providing important information, which are interpreted from observation as they cannot be analysed or determined from numbers (Dix *et al*., 1998: 437)). However, such a convenient distinction between quantitative and qualitative approaches can be blurred as some objective measurements can lend themselves to subjective analyses whilst conversely, it is sometimes necessary to quantify qualitative data using categorisation techniques or quantitative scales. This is particularly the case with questionnaires and checklists, which seek to draw statistics from qualitative information or subjective analyses from numerical scales.

2.1.4 User-centred evaluation methods

Due to space constraint, relevant methods will only be categorised according to their main characteristics and presented in a table format (see Table 1). This list provides a summary of the main user-centred evaluation methods on the basis of their appropriateness established from previous experience in CALL and online CALL evaluation. Refer to Dix *et al*. (1998); Dumas & Redish (1999) and Shneiderman (1998) for useful introductions and overviews, Boyle (1997) and Preece *et al*. (1994, 2002) for practical descriptions, detailed characteristics and comprehensive analyses, Karat (1997) and Preece *et al*. (2002) for the elaboration of evaluative frameworks and Ravden and Johnson (1989) as a particularly good example of a practical usability method. More specifically, refer to Hémard (1999, 2003) and Squires & Preece (1999) for methods such as user walkthroughs and heuristic evaluation applied to language learning and educational contexts respectively.

2.1.5 Limitations

As Levy (1997:41) pointed out, evaluation in the 1990s became particularly complex with the rapid introduction of new technologies, the range and interactive potential of new activities and the restricted role it played within the design process due to a lack of time and funding. This is still true today, although this central evaluation issue is fur-

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

Table 1. *Evaluation in online CALL*

Methods	When to use	How to use	Type of data	Advantages	Disadvantages
Questionnaire	Formative and summative	Best when contextualised and focused	Qualitative and quantitative	Expedient, cheap, can reach large numbers of users quickly.	Wording can be difficult, Students are notorious for not filling them in, so they are not always representative. Difficult to analyse accurately / efficiently. Should not be used on their own.
Informal feedback	Formative and summative	In class, following CALL interaction	Qualitative	Quick, easy, flexible, can produce some well focused feedback.	Requires availability and motivation from both evaluators and learners. Data can be parochial and misleading.
Real life observation	Summative	In class during a CALL activity	Qualitative	Can be a real eye opener on student interaction and learning experience.	Complements questionnaires well. Requires availability and observation can be difficult to interpret and generalise.
Checklist	Predictive and summative	Input from peers and expert knowledge	Qualitative and quantitative	Can generate valid feedback from peer group. Practical and cost-effective way of exploiting expert help. Useful to evaluate predictively (i.e. purchase of software, see Jones *et al.*, 1999).	Limitations in scope and use. Can be a mere box-ticking exercise rubber stamping a design decision.

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

Table 1. *Continued*

Methods	When to use	How to use	Type of data	Advantages	Disadvantages
User walkthrough (see for instance Lewis & Wharton, 1997)	Formative and summative	CALL lab with data capture equipment (video or tape)	Qualitative mainly, but can also be quantitative	Genuine recording of learner in context of use. Can generate good interaction between students (when paired) and computer. Produces large amount of genuine data.	Exploitation and analysis of data is time consuming. Can have resource implications and may require technical support for the video capture of interaction.
Focus group	Formative and summative	Planned dedicated space	Qualitative	Can generate fruitful discussions and provide focused feedback.	Can be problematic to set up with a busy or volatile student population. Requires skills related to approach.
Tracking	Formative and summative	CALL lab with interaction logging facility	Qualitative and quantitative	Help analyse users' behaviour and understand how tasks are carried out with unobtrusive software with synchronised visual aid.	Students often resent tracking devices (ref. to Big Brother) corrupting data. Very time consuming to analyse data unless analysis tool available. Feedback limited.
Usability testing	Summative	CALL lab	Quantitative	Valuable and accurate feedback when successful measure of users' performance to specific tasks obtained	Students can be unruly participants within a formal context like this.

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

Table 1. *Continued*

Methods	When to use	How to use	Type of data	Advantages	Disadvantages
Heuristic evaluation	Formative and summative	See heuristics in Nielsen (2000, 2001); Preece *et al*. (2002)	Qualitative	Straightforward application of expert guidelines, useful for critical appraisal of existing software, can generate reflective discussions and feedback from other colleagues.	Presents similar limitations to that of the checklist above, can be time-consuming and has potential resource implications (software) if done properly.

ther exacerbated by the more widespread involvement of mainstream practitioners often under the instruction of managers "nervous about innovations that may fail" (Boyle, 1997:201), the even greater range of devices and the seeming success of online CALL delivery, 'safely' behaving within recognised and approved web-based standards and practices. Evidence from the CALL literature suggests that if online CALL projects generally include varying forms of evaluative assessment, these tend to be more oriented towards the validation of an end-product rather than a means to evaluate an interface or its interaction, and are often presented as finite. In this sense, evaluation in online CALL in particular is still very much perceived as a discrete and often final stage in the design process, if time and resources permit it, whose outcome, given the context and constraints, is a predicted foregone conclusion. Students, who as end-users are rarely consulted about their own requirements (Hémard, 2003), are regularly quoted as saying via the medium of convenient questionnaires that they liked very much what they saw and did, thus giving new interfaces the necessary credibility and official seal of approval. Even if true, although many corrupting factors such as novelty, expediency of method and volatility of students come into play, there is an important need to understand the role of evaluation better in order to generate relevant and genuine data on design but also, crucially so, on its implementation. Whilst it is true that CALL practitioners, as front-users of authoring tools with pre-designed interfaces and functionality, can only play a limited part in the design, much benefit would be gained if they could systematically evaluate their design objectives as well as the impact their new interactive language learning environments had on their identified target learners.

3 Case study: usability in the context of use

3.1 Background

Much information has already been published on the London Guildhall University CALL project (for further details refer to Hémard & Cushion, 2000, 2001, 2002, 2003; Cushion & Hémard, 2003). Broadly, its original aims were to design and develop an online text-based and multimedia authoring tool in order to create and integrate a web-based interactive CALL environment into new and existing language curricula. Based on language teaching, Java programming and user interface design expertise, its objectives were to develop and reach a critical mass of material and evaluate its usability and utility within the university's language student population. The following is a brief account of its development over the last five years.

The CALL Project, initially entitled the *Web-based Language Learning Project*, started in 1998. The first year saw the development of a new authoring tool, which became instrumental in creating and introducing a new web-based environment for identified Level 1 French units. The subsequent evaluation throughout the second year led to version 1.2 of the authoring tool, still providing a text-based interaction over a range of different exercises. At the same time, Level 1 German and Spanish material started being inputted into the expanding online interactive language environment. In its third year of existence, an improved version 2, including beta sound authoring, was fully evaluated. This process led to the development of version 3 in the fourth year,

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

which combined full integration of sound and text authoring as well as online assessment such as listening comprehension, gapfill and multiple choice exercises. In parallel with this software development available for French, German, Spanish, Italian, Portuguese and English, the authoring tool was extended to Arabic, Tamil, and Punjabi. Now in its fifth year, the authoring tool has been renamed the "Guildhall Interactive Software for Multimedia Online (GISMO)" and can now be downloaded free from its new website: http://languages.londonmet.ac.uk/CALL/.

3.2 Interactivity

The original interaction was essentially grammar-oriented, complementing and exemplifying grammatical explanations whilst being predicated upon the range and type of interactive exercises generated by the authoring tool. Its integration into language programmes at Level 1 was based on the existing weekly progression and relied on a close match with the language material currently used. Although already successful, the introduction of sound considerably improved the interaction, increasing its level of acceptability by teachers and students alike. Likewise, to improve access and use by a wider student population, the online CALL material was redesigned as a grammar database, emphasising topics rather than progression, alongside a more open-ended exploratory learning environment with the addition of authentic texts and hypertext links to relevant websites. Thus the new design model (Norman, 1988) combining usability with a constructivist view of language learning, ensured that students could access support material, engage in and practise with grammar-based interactivities and navigate through related, theme-based, multimedia material on the web. Finally, to complete this evolutionary cycle, online assessment in the form of 'client-side' self-tests and formal 'server-side' tests (Arneil & Holmes, 2003; Hémard & Cushion, 2002) was introduced to counter-balance the largely score-free interactive environment and highlight the necessary learning output against the existing online interactive input.

However, despite the systematic evaluation of the interface's evolving usability, still little evidence existed to suggest that, beyond the behaviour of the system, the students' mental model of how the interactive environment worked matched our design model. The only information specifically related to the students' interaction in their context of use, originated from a previous evaluation conducted in 2000 (Hémard & Cushion, 2001) and anecdotal evidence from regular observations and was either in need of being updated or largely inconclusive.

3.3 Objectives of the case study

The case study's main objectives were twofold. It was initially designed to generate evaluative data on the students' interaction within their context of use, using different evaluation methods, as well as to highlight the case for evaluation by focusing on the chosen methods through a comparative analysis of their input and evaluative output. In particular, answers to the following questions were sought:

- How do students use the web? In other words, how conversant and literate are students with the web.

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

- Can students develop their own structural mental model (Allen, 1997; Preece *et al.* 1994) to match the design model?
- Do clear demarcation lines exist between language and cultural content?
- Does a multimedia approach provide added value?
- Are evaluation methods more efficient and accurate when appropriately chosen, combined and sequenced?
- Can complementary evaluation methods facilitate student engagement and help improve the quality of the feedback?

3.3.1 Structural model

Based on a theme, in this case France's recent colonial history, chosen for its relevance to an identified syllabus, a specific structural model was designed to exploit both the web's hypertext architecture as well as its modularity. As such, it sought to contextualise specific grammatical points with lexical and cultural aspects of the language by combining textual material, purposely fragmented into a tree structure, with a variety of relevant exercises and related multimedia material. Thus the main theme was subdivided into three sections: *Les Zoos Humains, l'Exposition Coloniale Universelle de 1931 à Paris* and *L'Ampleur du Colonialisme* which, in turn were further divided into three clearly focused texts. For example, *Les Zoos Humains* was dealt with from three different angles: *Le Contexte historique*, *Le Jardin anthropo-zoologique* and *La Hiérarchie des civilisations*. These manageable texts formed the basis for four different types of language-oriented interactivities: *Explication du vocabulaire*, *Lecture du texte*, an appropriate grammar exercise (e.g. gapfill exercise on past tenses) and an oral comprehension exercise (see Figure 1).

To complete the modularised process and to consolidate the interaction, additional multimedia support, such as maps and pictures and a number of self-tests, related to vocabulary and grammar, were integrated. Ultimately, the theme itself was presented as a link entitled *Le Temps des Colonies* within the grammar database under the past tenses (see Figure 1 which highlights the potential interaction linked to one branch of the structure).

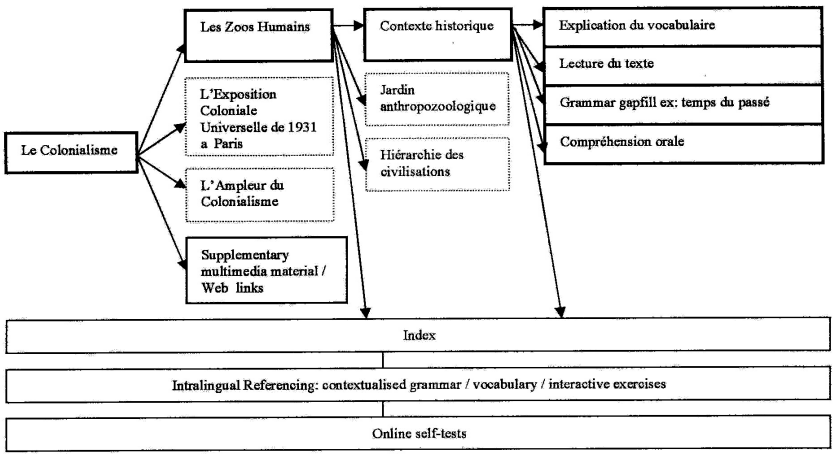


Fig. 1. Structural model.

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

Thus, the structural model, by combining specific cultural content with related language interactivity, was intended to be used to study the students' behaviour across different interfaces through their interaction, their navigation between grammar and content-related material as well as their different conceptions of learning (Goodyear *et al*., 2003).

3.4 Evaluative approach

Having established the purpose of the evaluation and what data needed to be collected, three methods were chosen on the basis of their complementarity but also existing constraints such as staff and student availability as well as limited resources. Firstly, a questionnaire, which sought ICT and CALL specific feedback, based on a Likert scale, was handed out to all fifty first year French Students at London Metropolitan University. Questions related to linguistic level, ICT experience, expertise and confidence, opinions on their interaction with the interactive language learning environment, its learning objectives, outcomes and assessment. The questionnaire was considered important since it was designed to update the previous student ICT profile and obtain information related to the students' CALL experience. Furthermore, it was also designed to facilitate the identification of a representative sample of six students, based on their level of French, their programme of studies and their ICT competence, who would be willing to volunteer for the following evaluation sessions. Secondly, four 2-hour user walkthrough sessions for three groups of two students each were organised. These were designed to observe the student interaction in three specific interactive contexts: the grammar exercises, the additional textual material and the new structural model (see detailed progression below). Finally, two 2-hour focus group meetings were planned: one as a fifth and final session with the same students based on their interactive experience during the user walkthrough sessions and one with a small group of three students who had not taken part in the user walkthroughs. Overall, this evaluation was based on sixty student hours and conducted by one member of staff. The user walkthroughs were planned as follows:

- Session 1: The beginning of the session was devoted to explaining the objectives of the evaluation and the chosen methods. The students were then asked to fill in the same questionnaire, this time following an explanation and a discussion related to its questions.
- Session 1: User Walkthrough (UW)1: free interaction with language learning environment in order to observe student strategies for using online CALL (grammar-based/activity-based/theme-based).
- Session 2: UW2: focus on related text-based activities and hypermedia links. Students were specifically asked to interact with and comment on the supplementary web-based material and how this interaction complemented grammar exercises.
- Session 3: UW3: task-based interaction: what can be learned from *Le Temps des Colonies* using the information provided, the support material and the interactive exercises? Observation of student strategies.
- Session 4: UW4: test-based interaction: how different are the students' approach and the learning objectives in *Le Temps des Colonies* when preparing for the

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

Table 2. *Comparison of methods: input*

Involvement	Questionnaires	User walkthroughs	Focus groups
Staff	Design, photocopying, distribution (50 students, 4 groups), collection, data analysis	Planning, organisation (logistics: location, computer,video-recording equipment), induction, transcription, data analysis.	Planning, organisation (logistics: location, computer, video-recording equipment), convening, transcription, data analysis.
Staff time	Approximately 10 hours	Approximately 60 hours	Approximately 20 hours
Students	Minimal: average 5 minutes	Real time: 4 x 2 hours	Real time: 2 x 2 hours
Resources	Minimal / negligible	Students were paid £5 per hour for both user walkthroughs and focus group discussions	

online self-tests?

- Session 5: Focus group: recapitulation/discussion generated by experience/ opportunity for the students to exchange views and impressions based on interaction, input and output.

4 Comparative analysis

4.1 Input

As can be seen from Table 2, the least onerous method in terms of staff time and resources was, unsurprisingly, the questionnaire. Whereas the user walkthrough, with its student involvement and huge volumes of data to present and analyse, was the most time-consuming and costly approach.

4.2 Output

The comparison between evaluative outputs was based on the representative nature of the resulting data, its relevance in relation to the objectives of the evaluation, the quality of the information and output variations depending on contexts of use (Table 3). Questionnaires are often deemed to generate the most representative and applicable data due to their outreach capability. However, the quality of responses due to the students' suspected lack of motivation or time and subsequent value of the ensuing analysis suggest that, by themselves, the "evidence of wide general opinion" (Preece *et al.*, 2002: 398) was, in this case, flawed and open to differing interpretations. Conversely, user walkthroughs and focus groups, relying on a genuine and dedicated student participation were deemed to be more valid and appropriate methods casting further light on and deepening our understanding of the student's interactive and learning experience.

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

However, since they mainly target small numbers of students at any one time, the data they generated could not be accurately generalised. Interestingly, all the methods used benefited from being combined together and fully integrated into the evaluation process.

5 Main findings: a representative selection

5.1 From questionnaires

- The overwhelming majority of students (95%) have experience in the use of computers.
- Of these, all use the Internet (most common use: email, music, general information), 75% feel confident but only 65% think they have the necessary expertise.
- Half the students claimed they use CALL in class. Of these, the vast majority (82%) claimed that they found the exercises useful, that CALL met their needs (78%), that they enjoyed practising with CALL (83%), that they liked the design and approach (80%) and that online tests were an incentive for practising with CALL (87%).
- Two specific comments were made: one student felt that CALL was too basic, too simple and not accurate enough. Another thought that CALL helped him/her to become more accurate in French and improved learning.

5.2 From user walkthroughs

Interface design issues

- The addition of pictures was welcomed generally but some students felt that, without supplementary information, these pictures were of limited appeal and use. This similarly applied to maps.
- The length of texts in selected websites was considered significant since the longer the text was, the more scrolling would ensue and the less productive the interaction became (see interactivity below).
- The lack of specific or targeted feedback was considered frustrating especially when long sentences needed to be entered in listening and gapfill exercises.

Interaction

- Much of the grammar-based interaction was mechanical and linear in approach. Typically, students would choose a topic, generally connected to their current syllabus, engage in the first interactivity and complete the range of exercises provided methodically. Interestingly, none of the students, with a few rare exceptions, clicked on a grammar topic without prior mutual agreement or on exercises randomly in the list provided.
- There was little evidence of interaction between the grammar database and the interactive exercises amongst the students apart from one group of intermediate students with low to average ICT confidence who made a few attempts at linking an exercise with grammar explanations.
- Disorientation only occurred as a result of interference between the browser's navigation functions and the exercises' own interface and only affected the more

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

Table 3. *Comparison of methods: output*

Data Output	Questionnaires	User walkthroughs	Focus groups
Representativeness	30 returned & filled questionnaires: 60%	Limited to a small but representative sample of six students.	Same as user walkthroughs.
Relevance to objectives	Student profiling: ICT use, confidence, competence, CALL practice and experience	Interface design issues student interaction, learning strategies, CALL practice and experience.	ICT use, confidence and expertise, CALL practice and experience learning strategies and to a lesser extent interface design issues.
Quality of data	Broad stroke evaluation yielding quantitative but impressionistic data:essentially a box-ticking exercise, 2 added written comments only.	Large amounts of directed qualitative data from analysis of students' natural, often revealing exchanges and real time interaction	Varied between the 2 groups. See contrast below. Rich in targeted qualitative data encompassing the students' experience and practice, especially with students who had previously taken part in the user walkthroughs.
Output variations	When the students were led through the questionnaire during the 1st user walkthrough, all answers were qualified by precise and relevant written comments.	Combining the questionnaire with the user walkthrough helps students to focus on areas of interest and concern. The questionnaire in this context was also instrumental in forming a basis for initial exchanges between the students	The two focus groups produced different results Students who had taken part in the previous user walkthroughs found it easy to comment on ICT and CALL, based on their shared experience. Conversely, students who had not taken part in the evaluation were far less voluble and, as a result, their focus group was more artificial. Thus, discussions needed a greater input and stimulus from the evaluator.

ICT literate students naturally conversant with the browser's functionality.

- Text-based interaction within the selected websites was largely unproductive as it essentially revolved round scrolling and navigating between links interspersed with fleeting glances at pictures and cursory reading.
- Students only seemed to develop a mental model of the interactive environment

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

when an overall structure linking text and interaction was made obvious. This was particularly the case when, for example, students realised that a range of different exercises were all related to a specific text. When this happened, greater attention was paid to the text and its meaning leading, in a number of cases, to an inter-exercise interaction such as looking for the meaning of a word previously seen or listening to a sentence again. This noticeable change of student engagement suggested a greater interest in a recurrent topic and a better understanding of the structure.

- Of greatest interest was the fact that the students reacted well to the structure of the *Temps des Colonies* design model. In contrast with their previous inability to relate and link complementary but juxtaposed material together, the majority of students engaged in the chosen subject matter and found a logical and motivating thread through its modularity. In this respect, this was the closest evidence as yet that mental models developed by students of how the system worked could match that of the designer's interactive structure.

Language learning issues

- There was a close correlation between the students' interactivity and their language syllabus. Students would often recall what they had studied to support their choice of exercises.
- Exercises, irrelevant of their length, were almost systematically completed, although wrong answers were not always corrected, multiple choice questions did not always engage students and few checked a grammatical point when completing an exercise. The longer exercises were criticised for their length.
- As previously mentioned, progression was mainly linear, logical and systematic when progressing through grammar exercises.
- Students also commented on the quality and relevance of texts, suggesting that a text had to be interesting in its own right, be authentic and combine culture, history and grammar. They particularly liked the same text, hence the same vocabulary, being used in different types of exercises.
- The three self-tests were undertaken thoroughly. They generated much student involvement and recorded their lengthiest concentration. This is partly due to the fact that they were linked to the theme, which had particularly attracted them and combined a listening comprehension with grammatical knowledge.
- Of particular interest, engagement had little to do with the students' deeper or lighter approach to their French studies in relation to the importance French played in their programme but was far more influenced by the designed interactive model.

5.3 From focus groups

- Students had no apprehension about computers although a generally shared premonition prevailed that the PC when used would crash for some unexplained reasons.
- Usage was very focused and essentially task-based. Its potential is still surprisingly limited and largely untapped. The students felt and recognised that they could do a lot more with the available functionality but were generally satisfied

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

with what they knew and could do, mainly in Word and with a web browser.

- A majority of students recognised they had poor keyboard skills but were all 'mouse-literate': as a result, navigation on the web was not a problem and a majority had a basic understanding of hypertext linking.
- Online CALL: exploring (clicking and scrolling) was by far the dominant type of interactivity. Students generally agreed that they did little reading on screen. As such, the book was still favoured for learning although the succinct nature of the material on the web was appreciated, resorted to and often printed out.
- The variety of interactivities meant that online CALL was a good way to learn but the general feeling was that it was not the best one as it could not replace the teacher with authentic material such as newspapers or radio and TV recordings.
- Students welcomed the link between culture and language but would not necessarily make such a link if it were not integrated into the design.
- When asked, students best remembered the diagnostic online tests and the cultural content from interactive sessions.
- Their best interactivities were listening and comprehension exercises.
- On the question of multimedia, they thought that it was useful as long as it supported and complemented the learning material as well as related to ongoing studies. Otherwise, the consensus was that it could be artificial and perceived more like a distraction, although, according to the students, animations distracted men more than women.
- Scoring was considered good but not vital and frustrating because of accents. Scoring was thought to be motivating if rewarding otherwise it was felt it could be counter-productive. Generally, self-tests were approved of, especially for revision purposes.

6 Outcome: a summary

6.1 Analysis of findings and evaluation of methods

As expected, questionnaires provided an overall picture of students who had experience in the use of computers and, of course, the web, although this experience, interestingly, did not fully translate into ICT confidence and expertise. Equally interesting was the fact that only half the students who filled in the questionnaire had used CALL in class, suggesting poor attendance or irregular use or both coincidentally. A few students did not even know what CALL meant and did not fill in any of the CALL-related questions. Naturally, since those students who did use CALL liked the design and approach, the evaluation could have stopped at this stage, simply exploiting these quantitative results to reaffirm CALL's success and establish a new student profile and trends.

User walkthroughs provided a finer albeit magnified insight into the students' experience within different interactive contexts of use, which produced small scale but valuable feedback on varying levels of student engagement with grammar exercises, online texts, website navigation and a specially designed structural model. This suggested that, from passive reception and interaction, students could react to different stimulus and develop their own mental models of designed interactive structures. Also of interest, students were critical of additional multimedia input which did not seem to fulfil a purpose,

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

or textual material which was not adequately contextualised, highlighting the need to correlate the interactivity with an existing language syllabus and to better combine language with culture.

Finally, focus groups brought in a more personal touch to the previous two evaluative methods, enabling students to talk freely about their apprehensions and inadequacies such as poor typing skills and lack of knowledge concerning computers. Students generally agreed that their interactive engagement on the web was often superficial or minimal, resorting to clicking and scrolling, and still preferred books when it came to reading. In this sense, they were not opposed to the idea of linking language and culture but this had to be properly built-in to be seen to be useful. Also interesting was the fact that online CALL was perceived differently as students often used the interactive database for a purpose such as revision before a test or practising specific exercises as instructed, hence their more systematic, almost mechanical approach to it.

6.2 Resulting changes to be made to the interface

As a result of this evaluation, it was decided:

- to facilitate the interaction between the grammar and exercises by creating a grammar link, which would bring up specific explanations within the interactive space.
- to remedy disorientation by redesigning the navigation buttons to make them more obvious and intuitive.
- to help standardise authoring practices regarding the length of exercises by including specific design considerations and recommendations in the documentation.

7 Conclusion

This paper and case study attempted to demonstrate the relevance of evaluation by highlighting its role in the design process and how valid, genuine data can be realistically obtained by choosing, applying and combining complementary evaluation methods in order to promote CALL practice and research. Beyond its limited validating function, evaluation must be seen as a means to generate this much needed information on new and existing online CALL environments and their impact on students, be it at the level of design, interaction or conceptions of learning. Equally, such reliable evaluative data must be better shared within the CALL community, since most design projects, due to their localised nature and limited resources, can only deliver piecemeal evaluation results. It is only by dint of doing this that online CALL practitioners will be able to go beyond task-based instructions and be instrumental in developing a larger and more accurate picture of online CALL's potential and use. Indeed, through a better understanding and application of evaluation methods to online CALL, combined with a drive for greater validity through collaboration, research in CALL and CALL design will progress further, beyond its restricted practices and institutionalised margins.

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

References

Allen, R. B. (1997) Mental models and user models. In: Helander, M. G., Landauer, T. K. and Prabhu, P. V. (eds.), *Handbook of Human-Computer Interaction*. North-Holland: Elsevier, 49–65.

Arneil, S. and Holmes, M. (2003) Servers, clients, testing and teaching. In: Felix, U. (ed.), *Language Learning Online: Towards best practice*. Lisse: Swets & Zeitlinger, 59–80.

Bainbridge, L. (1990) Verbal protocol analysis. In: Wilson, J. R. and Corlett, E. N. (eds.), *Evaluation of Human Work: A Practical Ergonomics Methodology*. London: Taylor & Francis, 161–179.

Boyle, T. (1997) *Design for Multimedia Learning*. London: Prentice Hall.

Cushion, S. and Hémard, D. (2003) Designing a CALL package for Arabic while learning the language ab initio. *Computer Assisted Language Learning*, in press.

Dix, A., Finlay, J., Abowd, G. and Beale, R. (1998) *Human-Computer Interaction*. London: Prentice-Hall.

Dumas, J. S. and Redish, J. C. (1999) *A Practical Guide to Usability Testing*. Exeter: Intellect Books.

Godwin-Jones, R. (2003) Optimising web course design for language learners. In: Felix, U. (ed.), *Language Learning Online: Towards best practice*. Lisse: Swets & Zeitlinger, 43–56.

Goodyear, P., Asensio, M., Jones, C., Hodgson, V. and Steeples, C. (2003) Relationships between conceptions of learning, approaches to study and students' judgements about the value of their experiences of networked learning. *ALT-J*,**11**(1): 17–27.

Hémard, D. (1999) A methodology for designing student-centred hypermedia CALL. In: Debski, R. and Levy, M. (eds.), *WorldCALL: Global Perspectives on Computer-Assisted Language Learning*. Lisse: Swets & Zeitlinger, 215–228.

Hémard, D. (2003) *Language Learning Online: designing towards user acceptability*. In: Felix, U. (ed.), *Language Learning Online: Towards best practice*. Lisse: Swets & Zeitlinger, 21–43.

Hémard, D. and Cushion, S. (2000a) Authoring a web-enhanced interface for a new language-learning environment. *ALT-J*, **8**(1): 41–49.

Hémard, D. and Cushion, S. (2000b) From access to acceptability: exploiting the web to design a new CALL environment. *CALL* **13**(2): 1–16.

Hémard, D. and Cushion, S. (2001) Evaluation of a web-based language learning environment: the importance of a user-centred design approach for CALL. *ReCALL*, **13**(1): 129–142.

Hémard, D. and Cushion, S. (2002) Sound authoring on the web: meeting the users'needs. *CALL* **15**(3): 281–295.

Hémard, D. and Cushion, S. (2003) Design and evaluation of an online test: assessment conceived as a complementary CALL tool. *CALL* in press.

Jones, A., Scanlon, E., Tosunoglu, C., Morris, E., Ross, S., Butcher, P. and Greenberg, J. (1999) Contexts for evaluating educational software. *Interacting with Comput*. **11**: 499–516

Karat, J. (1997) User-centered software svaluation methodologies. In: Helander, M. G., Landauer, T. K. and Prabhu, P. V. (eds.), *Handbook of Human-Computer Interaction*. North-Holland: Elsevier, 689–704.

Levy, M. (1997) *Computer-Assisted Language Learning – Context and Conceptualization*. Oxford: Clarendon Press.

Levy, M. (1999) Design processes in CALL: integrating theory, research and evaluation. In: Cameron, K. (ed.), *Media Design and Applications*. Lisse: Swets & Zeitlinger, 83–107.

Lewis, C. and Wharton, C. (1997) Cognitive walkthroughs. In: Helander, M. G., Landauer, T. K. and Prabhu, P. V. (eds.), *Handbook of Human-Computer Interaction*. North-Holland: Elsevier, 717–733.

Nielsen, J. (2000) *Designing Web Usability*. Indianapolis, USA: New Riders Publishing.

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press

Nielsen, J. (2001) Ten Usability Heuristics. http://www.useit.com/papers/heuristic.

Norman, D. (1988) *The Design of Everyday Things*. New York: Basic Books.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S. and Carey, T. (1994) *Human-Computer Interaction*. Wokingham, UK: Addison-Wesley.

Preece, J., Rogers, Y. and Sharp, H. (2002) *Interaction Design: Beyond human-computer interaction*. New York: Wiley.

Ravden, S. and Johnson, G. (1989) *Evaluating Usability of Human-Computer Interfaces. A Practical Method*. Chichester: Ellis Horwood.

Shneiderman, B. (1998) *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, Wokingham, UK: Addison-Wesley.

Squires, D. (1999) Usability and educational software design. Special Issue. *Interacting with Computers* **11**: 463–466.

Squires, D. and Preece, J. (1999) Predicting quality in educational software: evaluating for learning, usability and the synergy between them. *Interacting with Computers* **11**: 467–483.

https://doi.org/10.1017/S0958344004001727 Published online by Cambridge University Press