ReCALL **16** (2): 396–415. © 2004 Cambridge University Press
DOI: 10.1017/S0958344004001028 Printed in the United Kingdom

Putting principles into practice

JOAN JAMIESON
Department of English, Northern Arizona University, Flagstaff, AZ 86011-6032, USA
(*e-mail joan.jamieson@nau.edu*)

CAROL A. CHAPELLE
Department of English, Iowa State University, 203 Ross Hall, Ames, IA 50011, USA
(*e-mail: carolc@iastate.edu*)

SHERRY PREISS
Pearson Longman ELT, 10 Bank Street, 9th Floor, White Plains, NY 10606, USA
(*e-mail: Sherry.Preiss@PearsonEd.com*)

Abstract

CALL evaluation might ideally draw on principles from fields such as second language acquisition, language pedagogy, instructional design, and testing and measurement in order to make judgments about criteria such as elaborated input, feedback, collaborative learning, authentic tasks, navigation, screen design, reliability, validity, impact, and practicality. In this study, a subset of criteria were used to evaluate the design of English as a second or foreign language (ESL/EFL) online courses and assessments, Longman English Online. This article illustrates how a set of principles suggested evaluation criteria which, in turn, suggested particular variables for the instructional design; these variables, again in turn, suggested potential operationalizations which could be implemented as task features in CALL materials. Results of the judgmental evaluation indicated that most of the criteria were met, although some better than others.

1 Introduction

Commercial publishers and academic institutions invest large amounts of resources when developing CALL learning materials which might positively affect the language learning experience of many learners throughout the world. In view of the efforts devoted to this enterprise and their potential effects, applied linguists need to take on the responsibility of evaluating such materials in a way that can inform future practices. The discussion of CALL evaluation in the professional literature attests to the complexity inherent in evaluation issues. In part because of differences between CALL and other forms of instruction, evaluation is both essential and difficult to accomplish. Chapelle

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

(2001) argues that despite the merits of checklists for evaluating CALL in particular contexts, the evaluation process needs to be elaborated through articulation of the principles and processes entailed. The purpose of this paper is to illustrate how a set of principles informed a series of processes in the evaluation of a commercial CALL product for teaching intermediate English as a second or foreign language (ESL/EFL).

This evaluation was focused on the materials themselves rather than on the tasks that the learners actually carried out as they were using the materials; in other words, the evaluation was judgmental, as defined by Chapelle (2001). Chapelle distinguished between judgmental evaluation and empirical evaluation. Judgmental evaluation is based on the logical analysis of a CALL activity (that is, the focus of the evaluation) by the evaluator. Empirical evaluation is based on the quantitative or qualitative analysis of a CALL activity through observed data which are summarized by the evaluator; these data may be elicited from an individual, but more often come from a group of people. Although ideally evaluation should include both judgmental and empirical analyses, this article focusses on judgmental analysis.

Chapelle (2001) described the following criteria that can be used to indicate CALL quality: language learning potential, meaning focus, learner fit, authenticity, impact, and practicality. Because the most important of these criteria is language learning potential, in this evaluation we treated language learning potential as a principle. Chapelle also described criteria that Bachman and Palmer (1996:19-37) used to indicate the usefulness of computer-assisted language tests: reliability, construct validity, authenticity, interactiveness, impact, and practicality. In this study, a set of five criteria were chosen for the evaluation. We analyzed the extent to which the CALL materials provided evidence of language learning potential in terms of three criteria: enhanced input, interaction, and production. We analyzed the extent to which the CALL assessments provided evidence of usefulness, or quality, in terms of two criteria: authenticity and construct validity.

In Section 2, we explain our framework for principle-based evaluation. This is followed by a brief description of the CALL materials which were evaluated. The next two sections provide a description and analysis of the degree to which evidence for language learning potential and test usefulness was found in the CALL materials. Results of our evaluation are then summarized.

2 Principle-based evaluation

Our evaluation begins by outlining the principles that guided the evaluation, as indicated in the first column in Table 1. Principle One states the primary purposes of CALL materials, providing the potential for language learning. Principal Two highlights the contribution assessment makes to CALL materials. Principle Three states that that the criteria used to conduct the analysis should be drawn from theory and practice in SLA and language assessment. This relationship is depicted by the arrow drawn from the third principle to the second column in Table 1. A *criterion* is defined as a standard on which a judgment can be made. Criteria that address language learning potential as defined by Chapelle (2001) center on the degree of opportunity for a beneficial focus on form. Focus on form as discussed in the research on instructed SLA can be prompted through instructional activities which (1) draw learners' attention to specific aspects of the linguistic input, (2) engage learners in interactions requiring their negotiation and

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

Table 1. *CALL evaluation principles and their associated criteria and variables*

Principles	Criteria	Variables	Examples of Operationalized Variables
1. Language learning potential should be the central concern when developing and evaluating CALL	Enhanced input	Input salience, modification, elaboration	Text highlighted on the screen
	Interaction	Between people, between a person and computer, within a person	Learner-learner communication tasks; grammar help available
	Production	Planning, correcting production, using help	Taking notes suggested; model for constructed responses
2. CALL should be evaluated in terms of the quality of assessment in addition to instruction	Authenticity	Domain sampling	Correspondence between tasks on test and in courseware
	Construct Validity	Score interpretation	Meaningfulness of scores
3. The criteria used to conduct the analysis should be drawn from theory and practice in SLA and language assessment	↑	↑	↑
4. Judgmental evaluation through defined variables is valuable for examining materials in a way that can improve future design			

co-construction of meaning, and (3) prompt learners to produce "comprehensible output." Criteria that address the quality of language tests describe the links between the test, the material on which it was based, and the meaningfulness of the results. Authenticity in this case refers to the degree of correspondence between the test tasks and the CALL tasks. Construct validity refers to the appropriateness of the interpretations that are made on the basic test scores.

Principle Four states that one value of a judgmental analysis is to use defined variables to examine materials; determining if or how these variables are operationalized gives us a systematic means to improve future CALL designs. This can be accomplished through a sequence of inferential links. Together, the principles motivate particular criteria against which we evaluate CALL materials. Variables serve as indicators of the criteria that are evaluated. The operationalizations are intended to serve as justifiable realizations of the

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

variables, the last link in a chain of inference beginning with the principles. The two arrows drawn from the fourth principle to the third and fourth columns in Table 1 are intended to point to the importance of defined variables and their operationalizations in the evaluation of the CALL materials. The degree to which the operationalized variables occur as features of the CALL materials will form the basis of our evaluation.

3 ESL/EFL CALL series

We examined CALL materials that we have helped to develop. These materials were created to teach ESL/EFL online to adults and young adults and they were named *Longman English Online* (*LEO*) at the time we worked with them. (The courseware is no longer delivered online, but rather in a CD format; it has been renamed *Longman English Interactive* – see http://www.longman.com/ae/multimedia/). This courseware consisted of four levels: *LEO 1*, *LEO 2*, *LEO 3*, and *LEO 4*. For our study we used *LEO 3*. The *LEO* courseware was created via a proprietary authoring and publishing system created for Pearson Education by a group of instructional designers, programmers, graphic artists, and English language teaching (ELT) professionals. The screen designs and navigation icons were all field tested before the course was implemented. Consequently, *LEO*'s presentation is clear, easy to use, and professional looking.

LEO 3 (Rost & Fuchs, 2003) is the intermediate level of this video-based multimedia, integrated skills program. Each of its units continues the story of a young journalist who is covering a story about a soccer scandal. The units not only contain listening comprehension practice, but also include practice with grammar, vocabulary, pronunciation, speaking, reading, research on the web, and writing. Each unit concludes with a quiz. There are longer tests at the end of every four units called "module tests," and one at the end of all twelve units, called "the end-of-level test." Of the three types of tests, the quizzes are considered to have the lowest stakes for the learner. The module tests and end-of-level test are considered to have somewhat higher stakes than the quizzes, but by definition all of these tests are considered "low stakes" tests (Davies *et al*., 1999:185).

4 Criteria for language learning potential

Three criteria for language learning potential, namely enhanced input, interaction, and linguistic production, were examined. These are not the only criteria one could choose to investigate; they are, however, criteria which have a relatively strong theoretical, pedagogical, and empirical base.

4.1 Enhanced input

The evaluation of *LEO 3* seeks to identify places in the program where "input enhancement" (Sharwood Smith, 1993) has been operationalized. Theory and research on SLA converge to suggest that the likelihood of learners acquiring linguistic input increases if their attention is drawn to linguistic features, using techniques such as marking salience, modification, or elaboration (Schmidt, 1990; Sharwood Smith, 1993; Robinson, 1995; Skehan, 1998). Table 2 summarizes these "enhanced input variables" and provides operationalizations of input enhancement in CALL materials. Each of these variables is described in more detail in the following sections.

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

Table 2. *Types of enhanced input expected to be beneficial to learners (adapted from Chapelle, 2003)*

Enhanced input variables	Operationalizations
Salience	Visually marking a grammatical form on the screen through highlighting or animation or phonologically through stress; repeating a grammatical form or lexical phrase.
Modification	Offering modifications to initially presented input to make it understandable to the learner through any means that gets at the meaning (e.g. images, L1 translation, L2 dictionary definitions, simplification).
Elaboration	Increasing the potential for understanding the input through addition of plausible, grammatical L2 elaborations to the original text (e.g. defining relative clauses).

4.1.1 Salience

Linguistic input can be salient to learners for many different reasons. Some ways in which input can be made salient is by highlighting particular forms in the input or repeating the target forms throughout the materials. One technique used in *LEO 3* to draw learners' attention to the grammar point of a lesson was to use different font sizes and color. In *LEO 3*, three levels of color were used. Orange was chosen to indicate the primary grammar focus; purple, the secondary focus and green, the tertiary focus. In the grammar explanation about the meaning and use of modal verbs in English illustrated in Figure 1, the key term "had better" is presented in bolded font; it is also displayed in a different color from the rest of the text.

In addition to the use of color, audio repetition of the text on the screen and animation

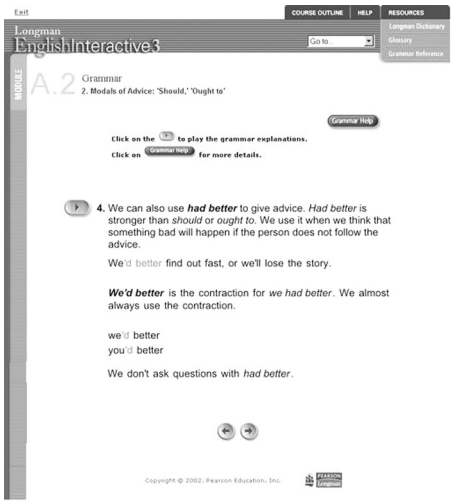


Fig. 1. Example of enhanced grammar input—salience.

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

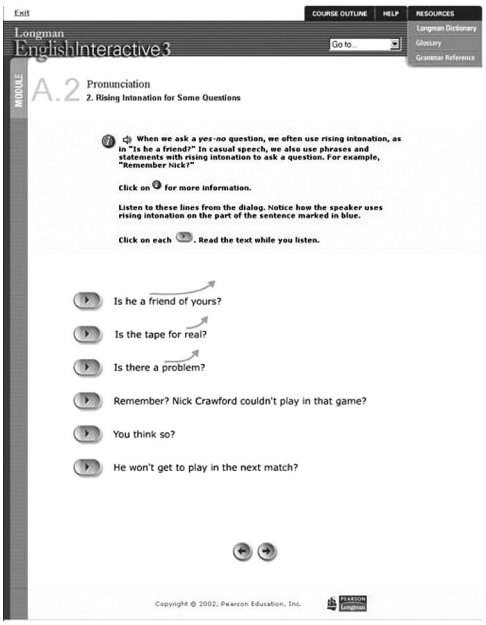


Fig. 2. Example of enhanced pronunciation input—salience.

were techniques used to draw the learners' attention. In this example, while the students listen to an audio presentation of the text that is on the screen, the words "had better" dissolve into "'d better." Choosing the appropriate color, font size, and animation required a tedious collaborative process among the ELT professionals, the technologist, and the production assistant skilled in working with the grammar authoring tool. In most cases the tool was able to represent the grammatical point desired by the ELT professionals, but there were cases in which the technology could simply not represent the desired features and "workarounds" that did not compromise the focus of the lesson had to be found.

Graphics and color were also used to reinforce the stress patterns which learners heard in the pronunciation portion of the lesson. In Figure 2, the rising intonation patterns of the questions are depicted. Note that the arrows are of different lengths.

Each of these arrows was specially designed by an outside design vendor in collaboration with an editorial team; consequently, they were very expensive and tedious to implement. Audio repetition, animation, and color changes, on the other hand, were comparatively inexpensive to include in the materials since they were achieved using a customized in-house tool. The use of color alone is a rather controversial technique due to the fact that many people suffer from color blindness, but using it in conjunction with other techniques increases the potential for salience on the part of many language learners.

4.1.2 Modification

Input modification refers to providing a modified or altered form of the linguistic input, typically in response to a signal of a of lack of comprehension. Based on research on face-to-face conversation in learning tasks, Larson-Freeman and Long (1991)

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

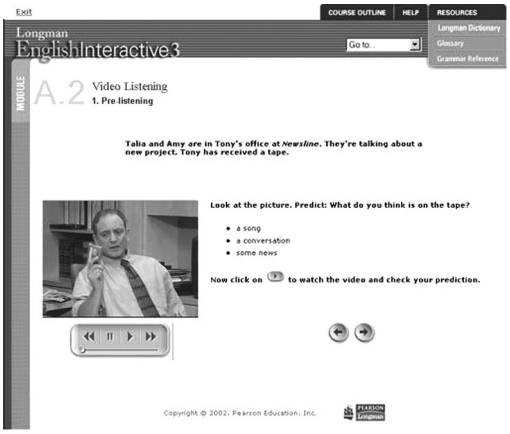


Fig. 3. Example of enhanced listening input—modification through images.

characterized modification as whatever an interlocutor does during the course of a conversation to clarify meaning to continue a conversation. Using this description, modifications can be any form of simplification, repetition, clarification, or first language (L1) translation. Extending this idea to CALL materials, one form of modification that gives learners access to the meaning of some vocabulary or other textual meaning is an image or a video depiction of the meaning.

The use of images and video are important aspects of *LEO 3*, as depicted in Figure 3. Here, we see the learner encouraged to look at the picture on the screen and to make a guess as to what material is on the tape. In every lesson, the learner not only listens to an ongoing story about the reporter, Talia, and the soccer star, Nick, but the learner watches a high-quality video of the story too.

Another form of modification that gives learners access to the meaning of what they are listening to is the use of transcripts. Figure 4 exemplifies how transcripts appear in

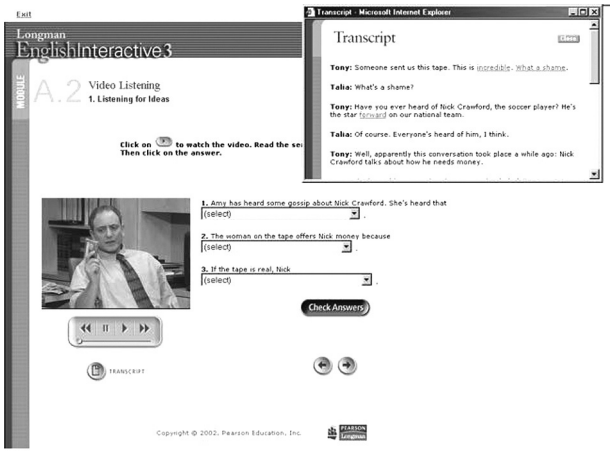


Fig. 4. Example of enhanced listening input—modification through transcripts.

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

the listening section, if the learner clicks on the transcript button.

Because the language used in the video was intended to portray authentic conversations, it contained colloquialisms and clipped pronunciation, and was delivered at natural speed. The transcripts provided another source of the L2 input for the learners, and it was hoped that they would listen and read to focus not only on the meaning, but also on the forms of the expressions that they heard.

4.1.3 Elaboration

Input elaboration is intended to help convey the meaning of the text by adding grammatical phrases and clauses such as defining appositives, relative clauses, and restatements (Yano *et al.,* 1993; Oh, 2001). In paper materials, researchers argue that elaboration is preferred over simplification. Simplification removes the forms that learners should be exposed to in the input. Elaboration adds to the input to clarify meaning while maintaining the structural and lexical complexity that learners need to learn from.

In the grammar explanation which is depicted in Figure 5, the "grammar coach" explains the usage of the present perfect in the indefinite past in English by incorporating several examples in her speech. These examples are also written on the blackboard. By adding these grammatical phrases to her explanation, her elaborated text provided more "comprehensible input" than might otherwise have been the case. The editors and designers had originally wanted the blackboard text to appear dynamically as the grammar coach spoke, but practical and technical considerations prevented that.

4.2 Interaction

Ellis (1999) describes the benefits of the second criterion for language learning potential, interaction, through three different theoretical perspectives: interaction hypothesis,

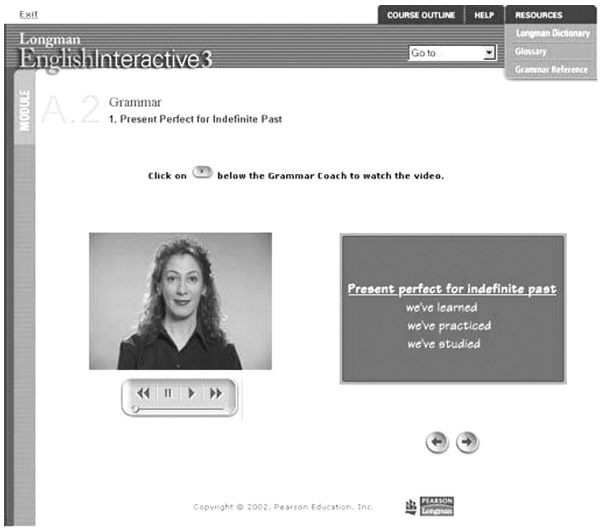


Fig. 5. Example of enhanced grammar input—elaboration.

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

sociocultural theory, and depth of processing theory. Each of these emphasizes different benefits of interaction, from which potential variables and operationalizations can be identified. The particular operationalizations depend on whether the interaction is "interpersonal," meaning "between people," "between a person and a computer," or "intrapersonal" (i.e. within a person's mind), as described in Chapelle (2003; see also Ellis, 1999; Hatch, 1978; Kol & Schcolnik, 2000; Lantolf & Appel, 1994; Long, 1996; Pica, 1994).

4.2.1 Interaction between people

Depending on one's particular theoretical orientation, interaction is hypothesized to benefit learners by providing opportunities for negotiation of meaning, co-construction of meaning, and prompting their attention to form. In *LEO 3*'s On the Web activities, learners were assigned to explore topics related to those in the video by going to preselected websites in search of answers to questions in the task directions. Afterward, learners had "web discussions" in which they chatted in groups of four addressing questions such as the following from Lesson 1:

> What are companies doing to keep their employees happy? Think about the web research you did and your own experience (or the experiences of other people you know). Discuss company benefits. Which benefits are most important to you? Why? (Johnson & Vaccara, 2003:5)

It was hoped that meaningful tasks with clear directions and a specific objective would encourage learners' participation in co-constructing responses to the questions. The Teacher's Guide suggested having each group choose four facts, and then write questions and answers for these facts. Next, each group would exchange its questions with another group, administering a sort of quiz. Members of the group would have to submit one set of answers to the group that sent them the questions. These are the types of directions that were intended to increase negotiation and co-construction of meaning. However, these directions did not encourage the learners to pay attention to form.

4.2.2 Interaction between a person and the computer

Interactions with the computer initiated through clicking on a hypertext link to receive help or by responding to a question can also be theorized through the three theoretical perspectives as beneficial because of the potential they provide for obtaining enhanced input and obtaining help for using language.

In Figure 6 we see three types of interaction between the learner and the computer that are possible in a listening comprehension exercise in *LEO 3*. The learner has selected answers to multiple choice comprehension questions and is shown graphically whether or not his/her answers are correct. If incorrect, the learner then has the option of clicking on "e" for an explanation (the part of the audio that contains the correct answer is replayed). Finally, the learner can click on "culture notes" which provide elaborated input for the video – in this example, the learner is told what the double-quotes indicated by finger movements mean in American English. The learner also has the option again of clicking on the transcript button. The transcript has key lexical items glossed, giving

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

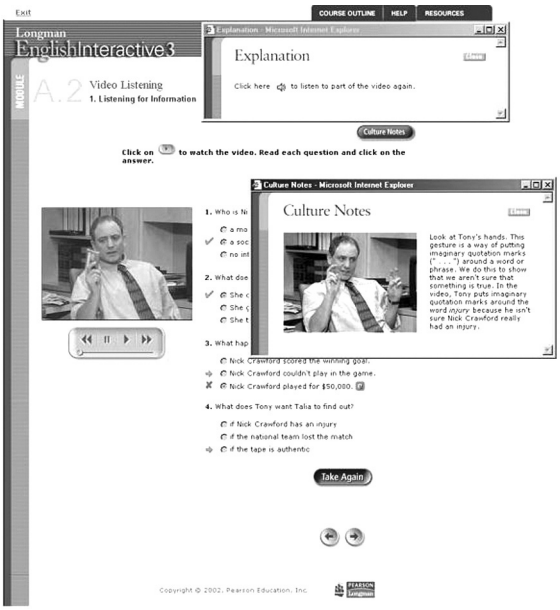


Fig. 6. Example of interaction between computer and learner.

the learner the choice of clicking on an unfamiliar word and bringing up a glossary that contains the definition, part of speech, and example sentence. Obviously, this type of interaction is a means of receiving some help with comprehension as well as negative feedback, and therefore two critical issues are (1) the quality of the interactions, and (2) the extent to which the learners actually engage in the interactions.

4.2.3 Interaction within a person's mind

Taking place in the learner's mind, intrapersonal interaction is hypothesized to hold value for prompting learners to focus on linguistic form, stimulating their inner voice, and getting them to engage in deep cognitive processing of input.

An example of how intrapersonal interaction plays out in *LEO 3* is shown in Figure 7. The screen shot demonstrates a game-like listening task which required extensive (and expensive) production and editorial work to achieve a sophisticated and engaging game-like language experience. In this listening comprehension exercise, the learner's task is to match events with dates in the career of a sports figure, Lance Armstrong. Intrapersonal interaction might be involved because the learner must decide how to approach task completion – all choices are self-implemented. When the learner has different paths available for task completion, he or she must take a more active role as a learner, and make conscious decisions, thus becoming more involved in the learning process.

4.3 Linguistic production

Theory and research on the third criterion for language learning potential, linguistic production, can also be conceptualized from several theoretical perspectives (He &

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

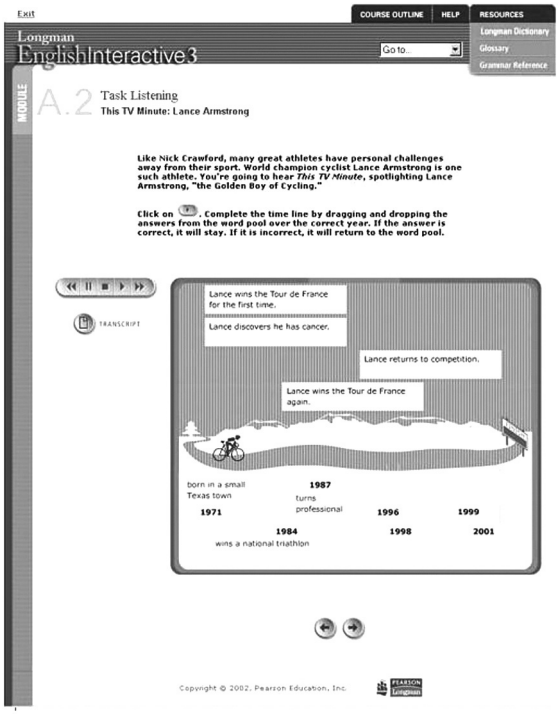


Fig. 7. Example of interaction within a person's mind.

Ellis, 1999; Nobuyoshi & Ellis, 1993; Pellettieri, 2000; Swain, 1985, 1998; Swain & Lapkin, 1995). From a cognitive or interactionist perspective, Swain (1985) pointed out that when learners produce "comprehensible output," they find out which lexicogrammatical forms they do not fully control. From a cognitive perspective, the benefits of producing language may be enhanced when learners have the opportunity to plan before speaking or writing and to correct linguistic output, which can be prompted by feedback from others or from self-evaluation. More recently, Swain (1998) has reconsidered the value of linguistic production from a sociocognitive perspective, focusing on the value of the help learners give and receive as they co-construct meaning during pair work. Such help from the interlocutor allows for production beyond what the learner can accomplish alone.

4.3.1 Planning

In computer-mediated communication tasks, particularly those using written language, learners can have the opportunity for planning before producing the language. In the pre-writing exercise depicted in Figure 8, learners are given a chart to fill in, as they search a website. This information will then be used in the paragraph they are assigned to write. Although the teacher's guides did offer suggested websites for the students, keeping the sites up to date was a challenge, but one that had to be tackled by the editorial team. In addition, finding sites that would be both general (not requiring a lot of background knowledge) and yet specific to the task was time-consuming.

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

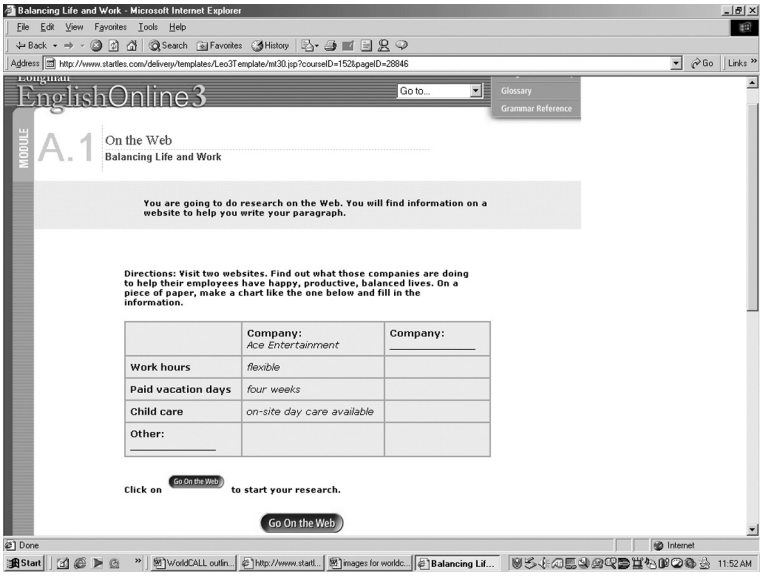


Fig. 8. Example of linguistic production—planning.

4.3.2 Correction

Correcting linguistic production refers to the learners' own correction of errors in the language they have produced. The recognition of the existence of errors may come from the learners' self-evaluation, from the signals received by other interlocutors or the teacher, or from computer-generated feedback.

In Figure 9 we see a speaking example from *LEO 3*. In this role-play, the learner speaks the lines of one of the two characters in a conversation. He or she can then compare the segment just spoken with the model in order to determine the degree of correspondence. Thus, we see an example of self-correcting here. The editors and developers were presented with three challenges: (1) how to make sure the students actually do listen to the model again, self-correct, and rerecord; (2) how to correct those students who do not follow the model but choose other correct and instructed choices; (3) how to design the speaking role-play in an intuitive, user-friendly manner. The first two challenges require instructor intervention and all three continue to remain problematic.

4.3.3 Help during production

Other types of help come from the preemptive guidance learners receive during linguistic production and the collaborative conversations in which they engage.

In *LEO 3*'s speaking and writing exercises, accurate online help was available to the learner in terms of glossaries, transcripts, grammar aids and models, and functional language boxes. Figure 10 depicts a speaking role-play in which the learner has selected to view both a functional language box and the transcript (in which highlighted words open

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

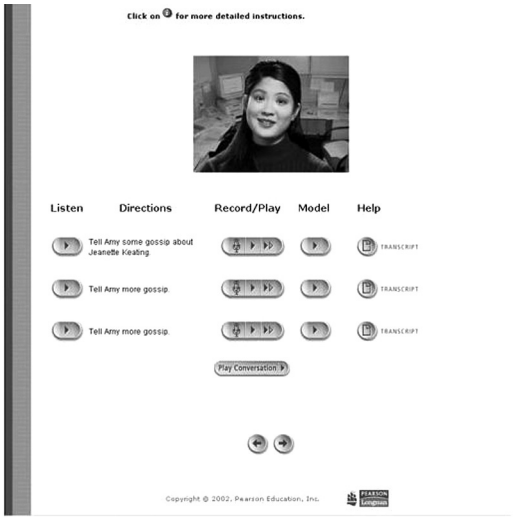


Fig. 9. Example of linguistic production—self-correction.

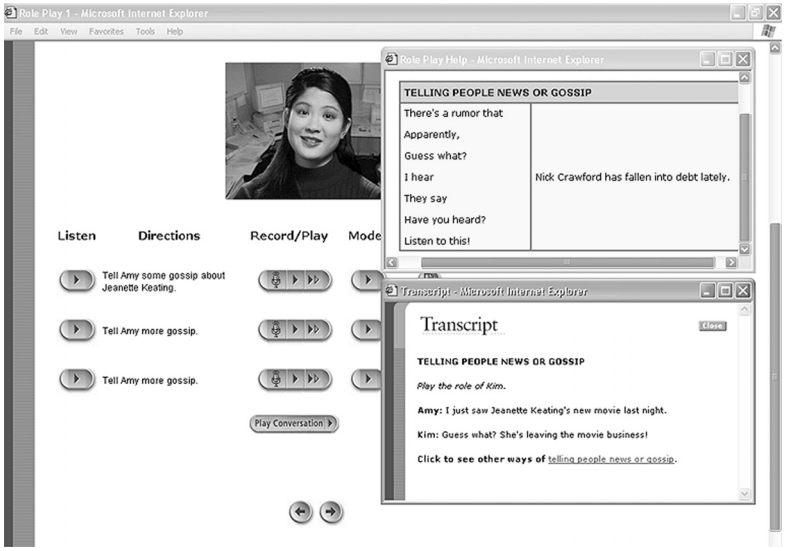


Fig. 10. Speaking role play with functional language box and transcript.

a vocabulary dialog box) that were presented in the unit.

5 Criteria for quality assessment

In this section, we turn our attention away from the instructional criteria for CALL tasks and explanations and toward the criteria for testing. The CALL tests in *LEO 3* were examined in terms of two criteria: authenticity and construct validity.

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

5.1 Authenticity

Bachman and Palmer (1996) described authenticity as the degree of correspondence between the characteristics of a test-task and the features of a task in "target language use" (TLU). They described the TLU in terms of real-life domains and language instruction domains. When the purpose of a test is to determine the degree of mastery of instructional objectives, the test is called an "achievement" test. In such cases, we are particularly concerned with the link between the test tasks and the tasks in the language instruction domain. An achievement test can be considered more or less authentic in terms of the degree to which its tasks and directions mirror the content and skills that were covered in an instructional unit (Davies *et al*., 1999). Mirroring the instructional content in a test is often referred to as *domain sampling* (e.g. Linn & Gronlund, 2000).

Domain sampling can be operationalized by covering the same knowledge, skills, or abilities from the instructional unit in a smaller, yet proportional number of tasks. Often, a developer creates a Table of Specifications to try and ensure proportional domain coverage. A Table of Specifications, in its simplest form, can be envisioned as a $3 \times n + 2$ matrix. The 3 columns across the top represent the skill areas, the number of items, and the percentage of items. There will be "$n + 2$" rows depending on the number of skill areas (n), followed again by (1) the number of items and (2) the percentage of items.

In Table 3, the design of the Table of Specifications for the *LEO 3* quizzes is shown. The skills that are tested in a quiz are listed in the first column. Their weightings (percentage allocations) are listed in the third column; these weightings sum to 100, as illustrated in the bottom-right-hand cell. These weightings are usually agreed upon by the test developers in consultation with the developers of the course. In this case, the majority of the course focused on listening and grammar (50%), followed by pronunciation and speaking (30%), and vocabulary and reading (20%).

Although the units in *LEO 3* contained On the Web research activities as well as some writing practice, it was decided not to include these skills in the tests because of the desire to limit the extent of online connections and to utilize machine scoring of learner responses rather than teacher scoring.

It was thought that forty items would be completable in between 30 and 40 minutes. Once the test developer knows the total number of items, that number needs to be multiplied by the proportional distribution for that skill. For example, 40 total

Table 3. *Table of specifications for LEO 3 quizzes*

Skill	# of items	% of items
Listening	10	25
Vocabulary	4	10
Speaking	6	15
Grammar	10	25
Pronunciation	6	15
Reading	4	10
# of items	40	
% of items		100

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

items × 0.25 (for Listening and Grammar) equals 10 items.

A second way to operationalize domain sampling is to cover similar content and item types as were in the instructional unit. In the case of *LEO 3*, only item types that had been used in the CALL tasks were used in the tests. Directions were the same for both tasks and test items. For all of the skills except listening, similar (but not the same) content to that used in the lessons was written by test developers, edited for both form and content, and rewritten. For listening, the same video that was used in the units was used in the tests, although it was edited into smaller segments. Ideally, video that was similar to that in the units, but not the same, would have been used, but this was not practical due to the costs of script writing, the hiring of actors, cameramen, developers, and editors.

5.2 Construct validity

Construct validity deals with the appropriateness and meaningfulness of the interpretations we make on the basis of a test score. There are a number of factors that come into play when discussing construct validity. We focused on the degree to which test scores were meaningful and appropriate. When using percentage correct as the method for score reporting, the question arises as to the minimum number of items that should be used. Referring back to the Table of Specifications, a general rule of thumb is that in order to adequately assess a learning outcome, there should be at least 10 items (Linn & Gronlund, 2000). In the 40-item quizzes, no subskill had more than 10 items, so only the total percent correct was reported to students as illustrated in Figure 11.

However, to make that score more meaningful, part scores were reported in terms of the number of items answered and the number of items which were correct; in this example, one of the authors got 4 out of 7 correct, listening for information items, and 2 out of 3 correct, listening for ideas items. An additional way to make the scores more

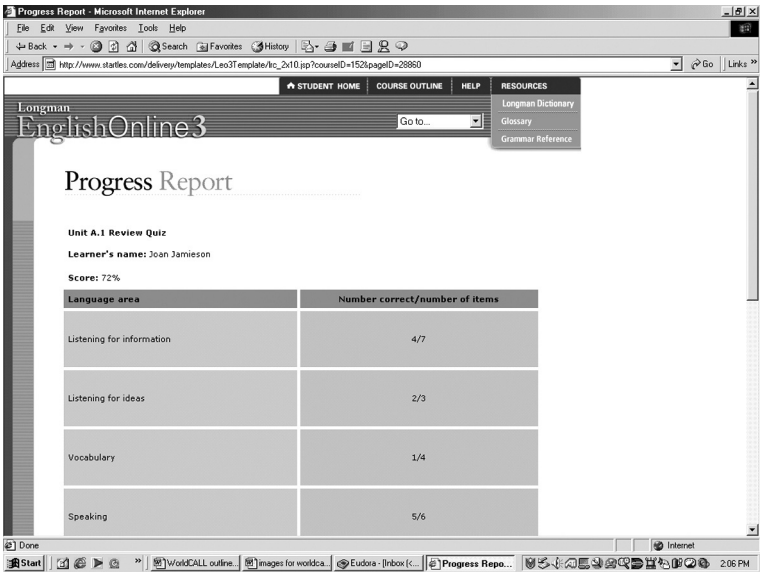


Fig. 11. Example of score reporting.

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press



Fig. 12. Example of performance information for Learner.

meaningful was to show learners what their answers were, whether they were correct, and if not, what the correct answer was. The learner just had to click on the skill area on the score report and then would go to the beginning of that section of the test.

In Figure 12 we see a section of the grammar quiz. Correct answers are given in green and incorrect answers appear in red with a strike though them. This was intended to let the learner think about his or her answer and compare it to the correct answer.

One final comment in this section on construct validity deals with the meaningfulness of the names of the skill areas on the tests. For the most part, the meaning of the skill on the test was the same as that in the unit, but this was not the case for speaking. Because of the desire to have only machine-scored items, the "speaking" items on the test involved only the recognition of appropriate words and phrases in given situations, such as inviting or making requests.

6 Summary

Table 4 summarizes the results of our judgmental evaluation of CALL materials based on criteria from theory and research. To evaluate the degree to which *LEO 3* successfully implemented the criteria and variables, we compared the evidence to the ideal operationalizations of language learning potential for the learning activities and of authenticity and construct validity for the assessments.

The developers' attempt to include enhanced input resulted in a good implementation of salience and modification, and a fair implementation of elaboration. Many examples of salience were found such as large font size and color, graphics, animation, and audio repetition. Modification was implemented through images and video, as well as through L2 definitions. Transcripts were used and offered as an example of modification.

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

Table 4. *Summary of Judgmental Evaluation of LEO 3*

Area	Criteria	Variables	Operationalizations	Evidence	Judgment
Language Learning Potential	Enhanced input	salience	font size/type/color, highlighting, animation, repetition	font size/ highlighting graphics, animation, audio repetition	Good
		modification	simplification, repetition, clarification, L1 translation, L2 definitions, image/video	images/video, transcripts, L2 definitions	Good
		elaboration	adding grammatical phrases or clauses to texts	adding grammatical phrases	Fair
	Interaction	between people	amount/types of co-constructed meaning, negotiations, and attention to form	meaningful tasks	Good
		people and computer	potential of providing enhanced input, obtaining help using the language	program correction of learner's response, explanations, culture notes	Good
		within person	potential for activating an inner voice	learner-initiated choices	Weak
	Production	planning	instructions to plan, directed activities, knowledge of expectations	directions to plan, pre-writing exercises	Good
		correcting	amount of self-correction, amount/ type of signals from computer or another person	self-correction of speaking/ pronunciation teacher correction	Fair
		help	accessing dictionary, grammar, models	glossaries, transcripts, grammar help, models of speaking and writing	Good
CALL Assessment	Authenticity	domain sampling	correspondence between tasks on test and in courseware	Table of specifications; directions, texts, item types	Fair
	Construct validity	score interpretation	appropriate, meaningful	reporting scheme	Good

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

L1 translation was not used in *LEO 3* (L1 translation is used for all directions in the lower-level *LEO 1* and *LEO 2*). Some evidence for elaboration was found for the grammar sections, but not for any of the other sections.

In terms of interaction, this judgmental analysis provided evidence that two of the three variables, interaction between people and between the person and the computer, were successfully implemented by the developers. On the Web tasks had clear directions for the teacher and a defined language task for the students, utilizing both synchronous chat and asynchronous discussion boards. Also, the use of program-generated correction, explanations of correct answers, and the variety of help options (vocabulary, grammar, culture notes, transcripts, glossaries) available to the learner provided good evidence that opportunities for person-computer interaction had been developed. Little evidence was found for opportunities for intrapersonal interaction.

Opportunities for learners' production in writing and speaking tasks were judged positively because their directions to the learners and the design of the activities involved planning. Both the speaking and writing tasks were intended to be handed to the teacher for correction. The directions encouraged learners to self-correct on the speaking tasks; however, the learners were not given directions about how the production would be evaluated by the teacher. Although it is possible that peer correction occurred during the discussions for the On the Web activities, it did not seem to be encouraged by the directions. There was no attempt to include signals from the computer regarding the correctness of linguistic output longer than a word or phrase. A lot of evidence for help was found in the material in the form of models, as well as transcripts, glossaries, and grammar help.

The authenticity of the tests was analyzed through the Table of Specifications. Although this showed good representation for many of the skills, it also revealed that the productive areas of speaking and writing as well as On the Web were not included in the tests. There was a high degree of correspondence between the materials and items on the test with the learning tasks in terms of item types and directions, but it was unfortunate that the same video and audio files were used in both tests and learning tasks.

Construct validity was judged on the basis of just one aspect of construct validity: the match between the intended construct interpretation and the score reporting. Scores were reported according to percentages or the number of items correct. The designers attempted to make the testing experience meaningful by allowing the learners to review test answers and to see correct answers. The naming of the skill areas matched the names of the sections of each unit such as listening, grammar, and vocabulary. These names are very broad, however, and they may mislead the learner into a false sense of confidence. For example, in the speaking section the learner had to recognize language functions, but never spoke.

7 Conclusion

The structure of this judgmental analysis allowed us to find evidence for enhanced input, interaction, planning, authenticity, and validity, though some evidence was more compelling than other evidence. Eleven variables were analyzed and "good" evidence was found in seven cases. A "fair" amount of evidence was found for three of the variables. In only one case was there "weak" evidence. Overall, based on our judgments on

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

the criteria of language learning potential as well as authenticity and construct validity of assessment, *LEO 3* received a good evaluation. This evaluation also points to areas in which *LEO 3* might be improved in the future and it suggests that *LEO 3* might be worth pursuing further through the addition of an empirical evaluation.

We acknowledge that this presentation has been incomplete on two counts. First, and foremost, we did not present both judgmental and empirical evidence for the criteria that could have been analyzed by both methods. This remains to be done, for as Chapelle (2001:59) wrote, empirical analysis of language learning potential asks the question: "What evidence suggests that the learner has acquired the target forms that have been focused on during CALL tasks?" Second, we have included only a subset of the criteria that we referred to from the literature. If we had examined meaning focus we would have addressed the cultural content of *LEO 3*. Interactiveness and practicality could also have been included but for lack of time and space. Last, but not least, we did not address situation or learner fit. In what context will these CALL materials be used? Are they appropriate for individual learners? Many questions remain.

As CALL designers and researchers, we seek new and less traditional approaches to evaluation. In this article, we suggest a model for CALL development and evaluation that takes into account principles which undergird a set of criteria which themselves are supported by theory and research in second language learning, teaching, and testing. If CALL developers become more familiar with criteria such as these, we can build upon them through implementation in our language learning tasks and tests. This, in turn, should result in CALL materials that have a sufficiently strong foundation that they will promote second language learning.

References

Bachman, L. and Palmer, A. (1996) *Language testing in practice*. Oxford: Oxford University Press.

Chapelle, C. A. (2001) *Computer applications in second language acquisition*. Cambridge: Cambridge University Press.

Chapelle, C. A. (2003) *English language learning and technology: lectures on teaching and research in the age of information and communication*. Amsterdam: John Benjamins.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. (1999) *Dictionary of language testing*. Cambridge: Cambridge University Press.

Ellis, R. (1999) *Learning a second language through interaction*. Amsterdam: John Benjamins.

Hatch, E. (1978) Discourse analysis and second language acquisition. In: Hatch, E. (ed.), *Second language acquisition: A book of readings*. Rowley, MA: Newbury House, pp. 401–475.

He, Q. and Ellis, R. (1999) Modified output and the acquisition of word meanings. In: Ellis, R. (ed.) *Learning a second language through interaction*. Amsterdam: John Benjamins, pp. 115–149.

Johnson, L. and Vaccara, R. (2003) *Longman English online communication companion*. New York: Pearson Education.

Kol, S. and Schcolnik, M. (2000) Enhancing screen reading strategies. *CALICO Journal*, **18** (1): 67–80.

Lantolf, J. and Appel, G. (1994) Theoretical framework: an introduction to Vygotskian approaches to second language research. In: Lantolf, J. P. and Appel, G. (eds.), V*ygotskian approaches to second language research*. Norwood, NJ: Ablex.

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press

Larson-Freeman, D. and Long, M. (1991) *An introduction to second language acquisition research*. London: Longman.

Linn, R. and Gronlund, N. (2000) *Measurement and assessment in teaching. 8th edition*. Upper Saddle River, NJ: Merrill, Prentice Hall.

Long, M. (1996) The role of linguistic environment in second language acquisition. In: Ritchie, W. C. and Bhatia, T. K. (eds.), *Handbook of second language acquisition*. San Diego, CA: Academic Press, pp. 413–468.

Nobuyoshi, J. and Ellis, R. (1993) Focused communication tasks. *English Language Teaching Journal*, **47**: 203–210.

Oh, S.-Y. (2001) Two types of input modification and ESL reading comprehension: simplification versus elaboration. *TESOL Quarterly*, **35** (1): 69–96.

Pelletieri, J. (2000) Negotiation in cyberspace: the role of chatting in the development of grammatical competence in the virtual foreign language classroom. In: Warschauer, M. and Kern, R. (eds.), *Network-based language teaching: Concepts and practice*. Cambridge: Cambridge University Press, pp. 59–86.

Pica, T. (1994) Research on negotiation: what does it reveal about second language learning conditions, processes, and outcomes? *Language Learning*, **44**: 493–527.

Robinson, P. (1995) Attention, memory, and the 'noticing' hypothesis. *Language Learning*, **45**: 285–331.

Rost, M. and Fuchs, M. (2003) *Longman English on-line 3*. New York: Pearson Education.

Schmidt, R. (1990) The role of consciousness in second language learning. *Applied Linguistics*, **11**: 129–158.

Sharwood Smith, M. (1993) Input enhancement in instructed SLA: theoretical bases. *Studies in Second Language Acquisition*, **15**: 165–179.

Skehan, P. (1998) *A cognitive approach to language learning*. Oxford: Oxford University Press.

Swain, M. (1985) Communicative competence: some roles of comprehensible input and comprehensible output in development. In: Gass, S. and Madden, C. (eds.), *Input in second language acquisition*. Rowley, MA: Newbury House, pp. 235–253.

Swain, M. (1998) Focus on form through conscious reflection. In: Doughty, C. and Williams, J. (eds.), *Focus on form in classroom second language acquisition*. Cambridge: Cambridge University Press, pp. 64–81.

Swain, M. and Lapkin, S. (1995) Problems in output and the cognitive processes they generate: a step towards second language learning. *Applied Linguistics*, **16**: 371–391.

Yano, Y., Long, M. and Ross, S. (1993) The effects of simplified and elaborated texts on foreign language reading comprehension. *Language Learning*, **43**, 297–308.

https://doi.org/10.1017/S0958344004001028 Published online by Cambridge University Press