Language Learning & Technology
http://llt.msu.edu/vol4num1/hegchap/default.html

May 2000, Volume 4, Number 1
pp. 39-55 online
(page numbers in PDF differ and should not be used for reference)

METHODOLOGICAL ISSUES IN RESEARCH ON LEARNER-COMPUTER INTERACTIONS IN CALL

Volker Hegelheimer & Carol A. Chapelle
Iowa State University

ABSTRACT

CALL materials may provide a mechanism for implementing theoretically-ideal conditions for second language acquisition and for conducting empirical research to investigate effects of these conditions. This paper explores methodological issues involved in realizing this potential by focusing on investigation of the *noticing hypothesis* (Schmidt, 1990) in CALL reading materials. It reviews the problem of assessing noticing in classroom and experimental settings through a) conditions for noticing, b) retrospective assessment, and c) concurrent assessment. Concurrent assessment, which provides the most direct measure of noticing, is illustrated through CALL materials that gather data on noticing, test retention of word meaning, and calculate the correlation between noticed and remembered words. Methodological issues of implementation and validation are discussed.

Teachers and researchers interested in improving the effectiveness of CALL activities sometimes look for guidance from second language acquisition (SLA) research with the hope that CALL activities can be designed to create ideal conditions for SLA. Despite convictions that SLA theory and research should inform CALL practice, details of how to form such links need to be spelled out. One might look to the SLA research that has used computer-assisted materials for developing experimental language learning tasks and for gathering performance data, but with some exceptions (e.g., Doughty, 1991; Hulstijn, 1993), the tasks used in such research appear decidedly experimental because they require participants to learn specific forms of an artificial language, for example. Even such experiments on learning rules of a natural language may require learning specific aspects of a language not of the learners' choosing for short duration determined by the researcher. Although such experiments carefully model the desired cognitive characteristics for formal learning, critical elements of learner motivation and communicative language use are likely to be missing. In fact, given the artificiality of the learning situation created by the laboratory experiment, Hulstijn warns that ". . .without additional research in real L2 learning environments, one should be extremely cautious in drawing immediate conclusions from laboratory studies to language pedagogy" (Hulstijn, 1997, p. 132).

This paper addresses the need for research in real L2 learning environments by illustrating how SLA theory can inform the design and evaluation of a CALL task that can be used in language classes or for self study. Methodological issues in the study of CALL from the perspective of interactionist SLA theory (Gass, 1997; Long, 1996; Pica, 1994) are explored through analysis of how one of its hypotheses, the noticing hypothesis (Schmidt, 1990, 1992), can guide the process of CALL development, data collection, analysis, and interpretation. The methodology is illustrated through sample texts in English, French, Spanish and German which allow readers to play the role of an L2 learner--who reads the text, requests definitions, and takes the post-test--and the researcher--who sees the report of student data and the statistical analysis. Discussion of the methodology highlights two issues: the quality of the task for creating conditions for noticing and the validity of assessment of noticing through this method.

SLA THEORY, CALL MATERIALS, AND EMPIRICAL RESEARCH

Copyright © 2000, ISSN 1094-3501

CALL developers and users constantly seek pedagogical principles that can guide their construction of CALL tasks to make them effective for SLA. At the same time, some SLA researchers seek to improve the ecological validity of studies by conducting them in settings where L2 learners are eager and motivated participants. CALL activities designed in view of hypotheses about ideal conditions for SLA can meet both objectives. This symbiotic relationship between CALL and SLA is maintained through a combination of theoretically-based hypotheses to inform the design of CALL materials and empirically-based research which evaluates both theory and materials (illustrated in Figure 1).

Figure 1. The relationship between SLA theory, CALL materials, and empirical research

Interactionist SLA Theory

Theory can be approached in a variety of ways, but for evaluating CALL materials, a theory needs to hypothesize characteristics of the linguistic environment that may be valuable for SLA (Doughty, 1987). Such a theory, interactionist theory, has been articulated primarily through a research program on the role of linguistic input and interaction in SLA in instructional settings (Gass, 1997; Long, 1996; Pica, 1994). Consequently, it makes hypotheses which are relevant to the design and study of CALL. First, interactionist theory claims that linguistic input, such as that received through CALL materials, needs to become *intake* in order to be acquired by the learner. Intake refers to input that the learner has comprehended both semantically and syntactically. Importantly, linguistic input that has been comprehended semantically may be of limited help to the learner because semantic comprehension is often accomplished by recognition of isolated lexical items or interpretation of non-linguistic cues with the help of existing schemata.

Second, input is more likely to become intake if it is noticed, and therefore Schmidt (1990) hypothesizes that noticing is necessary for acquisition. He claims ". . .that subliminal language learning is impossible, and that intake is what learners consciously notice. This requirement of noticing is meant to apply equally to all aspects of language (lexicon, phonology, grammatical form, pragmatics. . ." (Schmidt, 1990, p. 149). The noticing hypothesis is drawn from theoretical extensions from psychological research on attention and memory (Robinson, 1995; Schmidt, 1990) as well as the seminal introspective study documenting the role of attention and noticing for a beginner learning Portuguese through a combination of formal classroom instruction and immersion (Schmidt & Frota, 1986).

Third, learners are most likely to notice linguistic form during interaction. The most useful interactions are those which help learners comprehend the semantics and syntax of input and which help learners to improve the comprehensibility of their own linguistic output. Such beneficial interactions can occur in a number of different ways depending on the situation. In face-to-face conversation, comprehension can be achieved through negotiation of meaning which occurs during communication breakdowns when learners are confused about meaning or syntax and are therefore unable to comprehend the message at first. One reason that negotiation of meaning is valuable is that it can result in modified input--input which is better tuned to the learner's level of ability. Unlike intake and noticing, negotiation of meaning can be observed sequences of discourse including requests for repetition (e.g., Huh?), or clarification (e.g., What do you mean?). These and other observable interactions may indicate that learners have noticed the input.

CALL Materials

In pedagogical materials, principles for making intake from input through noticing have been introduced as the construct of *focus on form* (Long, 1988).

> *Focus on form refers* to how [the learner's] focal attentional resources are allocated. Although there are degrees of attention, and although attention to form and attention to meaning are not always mutually exclusive, during an otherwise meaning-focused [interaction], focus on form often consists of a shift of attention to linguistic code features--by the teacher and/or one or more students--triggered by perceived problems with comprehension or production. (Long & Robinson, 1998, p. 23)

Ideally, not only traditional classroom tasks, but CALL tasks as well, can provide conditions for learners to focus on form. Although some attempts at links between constructs of SLA theory and CALL design have been made (Chapelle, 1998; Doughty, 1987), much territory remains to be explored in the principled development and evaluation of CALL materials.

Evaluation requires investigation of the extent to which conditions for noticing can be shown to be related to retention of what was noticed. Schmidt (1992) described the problem as follows:

> It is. . .important to operationalize concepts such as intention, noticing, and awareness in both experimental and pedagogical settings, recognizing that the major issue in resolving problems of awareness and learning in psychology has been lack of consensus as to what constitutes an adequate measure of awareness. . . (p. 218)

Despite this difficulty, the effects of noticing have been investigated to some extent in several studies.

EMPIRICAL INVESTIGATION OF NOTICING

Research on noticing has taken three approaches to assessment. One method is to construct conditions for noticing in instructional or experimental materials by highlighting particular linguistic features. Learners who use those materials are then assumed to have noticed the target points. The second method asks learners to retrospectively report what they had noticed during task completion. The third method is to infer noticing from observable interactions such as negotiation of meaning during task completion. In none of these cases is noticing observed directly; each requires an inference, but they differ in the size of the inference that is made.

Inferences About Noticing From Conditions

Some researchers have attempted to investigate the effects of noticing through input enhancement (Sharwood Smith, 1993) conditions in which teachers or materials are expected to draw learners' attention to aspects of the input. For example, White (1998) provided francophone learners of English in grade 6 with texts in which particular grammatical forms were highlighted through use of a bold font. Three groups of students received 10 hours of instruction over a two-week period. The target structure--possessive determiners in French--was not visually enhanced for one group of students. Instead, the past tense marker —*ed* was enhanced to account for any effect text enhancement might have. Two groups received enhanced input (i.e., all third person singular pronouns and possessive determiners were italicized). Additionally, one group was exposed to more naturally occurring correct instances of the target form through a supplemental book. Using a baseline test, an immediate pre- and post-test, and a delayed posttest, White concluded that while the noticing prompted by enhancement may accelerate the acquisition of highlighted features, it may not be effective for L1-L2 contrasts such as the one researched in this study.

A study that found positive effects associated with input enhancement investigated the effects of highlighting relative clauses in texts presented in CALL materials. Doughty (1991) found that learners who used the materials with relative clauses highlighted in the texts outperformed another group receiving no highlighting when tested on their comprehension of the texts and their knowledge of relative clauses. Importantly, learners were instructed to read the text for meaning. Whereas all learners were assumed to focus on meaning, those provided with highlighting were assumed to have noticed the relative clauses in the text.

Without minimizing the contributions of this research design, researchers should note that this method of assessing noticing requires a high degree of inference because even though learners are presented with a condition in which noticing is prompted, it is not possible to assess exactly when and where noticing has occurred. Schmidt refers to this as an external approach to noticing, pointing out that it is incomplete for the investigation of noticing.

> The problem with this external approach is that the treatment may not have the intended effect. Learners may look for rules even when told not to, fail to notice highlighted words in the text, fail to understand grammar explanations, or fail to see how such explanations apply. . .so there is no choice but to continue to approach these questions from both external and internal perspectives, giving equal attention to experimental treatment and the assessment of learner awareness. (Schmidt, 1992, p. 219)

Retrospective Assessment of Noticing

Addressing the need for a more direct assessment of noticing, researchers have attempted to assess the extent to which learners have, in fact, noticed highlighted input by probing their knowledge of a structure after receiving input or examining retrospective think-alouds. A study investigating effects of input enhancement of preterit and imperfect verb forms in Spanish, for example, found that 14 second semester adults receiving enhanced input (underlined, bolded, and shadowed) made more references to these forms during think-alouds on a subsequent production task than learners who had not received input enhancement (Jourdenais, Ota, Stauffer, Boyson, & Doughty, 1995).

However, because the think aloud was conducted after learners had received the input the authors concluded that

> [t]he difference in mention of the target forms indicates that the enhancement participants had not been primed for processing of the target forms. This may have been caused by increased registration of the input stimulus due to textual modification. Whether most participants were aware of the enhancement, however, cannot be determined by this study, as only one enhancement subject remarked on the appearance of the sample text. (Jourdenais et al., 1995, p. 206)

In other words, this method requires a large inference to be made from learners' retrospective reports to their noticing during task completion.

Concurrent Assessment of Noticing in Oral Discourse

A number of classroom studies have investigated noticing during face-to-face oral conversation through observation of learners' conversational adjustments (Larsen-Freeman & Long, 1991), requests for modified input (Ellis, 1995), and language related episodes (Swain, 1998). The study of conversational adjustments signaled by clarification requests (e.g., What do you mean?) or requests for repetition (e.g., Huh?), for example, have become a mainstay in research on L2 learning tasks. Researchers believe that

conversational adjustments signal miscommunication which draws learners' attention to the language during otherwise meaning-focused activity (e.g., Gass & Madden, 1985).

Investigation of the effects of requests for modified input is based on the same general hypothesis–that communication breakdown allows for noticing which can result in comprehension if input is modified appropriately. One study investigating the relationship between modified input and retention of the words that had been modified found that premodified input promoted acquisition of word meanings (Ellis, 1995). Fifty-one third year High School students in Japan were asked to listen to directions and identify objects and place them in a specific position on a picture of a kitchen. The input was either premodified (i.e., by adding redundancy to the instructions) or interactionally modified (i.e., by allowing students to ask clarifying questions while the instructions were being read). Investigation of interactionally modified language was accomplished through concurrent assessment. In other words, learners requests for modified input were documented as they occurred.

Language,related episodes are observed when students engage in metatalk, or "talk about the language of the text they were reconstructing" (Swain, 1998, p. 71) during a dictogloss task requiring learners to collaboratively reconstruct a text. One study investing such tasks found that the approximately 40% of the instances of metatalk--language related episodes--focused on form while about 30% focused on word meaning. Both types of language-related episodes were hypothesized to be beneficial for acquisition because they were evidence that the learners were noticing aspects of the language rather than engaging in problem-free reconstruction. As a consequence, the research question asks to what extent the learner acquires the language that is the object of attention during the language-related episode, and language-related episodes are investigated through concurrent assessment to observe the precise object of the linguistic metatalk.

What conversational adjustments, requests for modified input, and language related episodes have in common is that they each indicate precisely when and where the learner stopped and noticed the language due to a problem in comprehension. Based on her concurrent assessment of language-related episodes, Swain (1998) argued for the importance of such concurrent assessment: "it seems essential in research to test what learners *actually do*, not what the research assumes instructions and task demands will lead learners to focus on" (p. 80) [italics in original]. She was referring particularly to the study of oral language; however, the same need exists for the assessment of noticing in written language. As a consequence, the following statement by Schmidt (1990) remains relevant today by calling for research designs which include concurrent assessment of noticing:

> There is almost a complete lack of evidence in the second language literature which is directly relevant to the [noticing] hypothesis, since second language researchers have never asked learners to provide systematic information on what they notice while learning languages that could be compared to what they can be shown (by other measures) to have learned. (p. 139)

ASSESSMENT OF NOTICING VOCABULARY IN READING

The majority of studies of noticing have focussed on grammatical form, but at least equally important to investigate is the role of noticing on vocabulary retention. The need for concurrent assessment may be even stronger in the study of vocabulary. When grammatical form is studied, researchers supply learners with multiple instances of a particular form or contrast that they are expected to notice in the input, and therefore their noticing at any point in the input may help to strengthen their knowledge of the form. Perhaps the more they notice the better, but the effects of noticing can be investigated through assessment of learners' improved knowledge of the target grammatical form. In contrast, vocabulary items in reading typically appear in the input once, and therefore investigation of noticing needs to record who noticed what. In a text with some unknown lexical items, the learners should be expected to notice vocabulary

which impedes their progress in reading for meaning. They then need to receive modified input as either L1 translations or L2 simplification to make the input intake, and therefore make it potential for acquisition.

Conditions

The value of conditions that provide modified input for vocabulary has been found in studies in which learners have access to glosses for difficult words during reading. For example, Watanabe (1997) found that EFL learners who were given access to marginal glosses retained previously unknown words better than did learners who were not offered marginal glosses with their readings. One might speculate that both groups noticed those words that they stumbled over as they were attempting to interpret the text, but that only those students in the condition with the gloss were able to recover from the breakdown through the use of the glosses.

Studies of vocabulary look up behavior relying on paper texts and glosses can only be implemented through a "condition" research design, unless obtrusive or impractical methods are used to observe behavior. In other words studies with texts and glosses on paper rely on what Schmidt refers to as the "external" approach rather than attempting to assess exactly when the learner noticed a particular linguistic form.

Concurrent Assessment of Noticing Vocabulary in Reading

Two approaches have been taken to address the need for concurrent assessment, both of which record learner's requests for modified input, during their on-line reading. The first summarizes the data to indicate who noticed, resulting in a classification of participants into those who requested glosses frequently and those who requested glosses infrequently.

A study that documented learners' look-up behaviors was conducted using second-year German students who were given on-line texts to read as part of their regular classroom assignment (Chun & Plass, 1996). As they read, they had access to annotations to words that the researchers believed would be difficult. The researchers recorded all requests by learners and conducted detailed analyses concerning the types of annotations (text, picture, or video) learners chose. The relationship between noticing and retention was examined by calculating the correlation of the number of look-ups each learner had made with their overall performance on the vocabulary test. Their conclusion fails to support the noticing hypothesis: "high frequencies of words looked up do not necessarily result in better performance on a vocabulary test" (p. 192).

Another study assessing noticing in the same way yielded contrary findings. Hegelheimer (1998) recorded individual instances of noticing and requests for input modification as expressed through mouse-clicks. After completing an online vocabulary pre-test, students read three short passages and answered nine questions about meaning following each passage. As the participants proceeded through the experiment, they had access to progressively more contextual help. The first passage provided students no access to help functions; they had access to textual glosses (definitions in the target language) in the second passage and access to textual glosses and sentence-level audio glosses (students could listen to the passage sentence-by-sentence) in the third passage. Results of a vocabulary post-test administered one week after the treatment indicated that the learners who frequently requested glosses (i.e., noticed and received modified input) performed significantly better on vocabulary post tests than those who used the glosses infrequently (Hegelheimer, 1998).

The second approach to concurrent assessment was taken in a study by Hsu (1994), who looked at the relationship between what was noticed and improvement in listening comprehension. Hsu documented noticing and receipt of modifications evident by learners' requests for repetitions, written transcriptions,

or written definitions for words in aural input in a CALL lesson. She also recorded the specific linguistic input associated with each of the learners' requests. She then assessed outcomes through pre- and post-tests which had been constructed specifically for the research to include the lexical phrases in the input. She found significant relationships between noticing and improvement in listening comprehension on lexical items for which modifications had been requested. Even though improved comprehension is only one facet of acquisition, and no delayed post-test results could argue that the effects lasted, this methodology provides an example of how concurrent assessment of noticing can be combined with assessment of improvement.

The findings from the latter two studies are consistent with the hypothesis from interactionist theory--that noticing unknown vocabulary and receipt of modified input increases chances for acquisition. As a consequence, this is a promising line of research to implement through CALL materials.

Assessing the Relationship Between Noticing and Retention

The following example provides an illustration of how the theoretical hypothesis about the value of noticing is operationalized through an on-line reading activity, and how the hypothesis is tested through collection and analysis of particular interaction data.

The CALL Materials

The CALL reading activity is intended to allow learners to acquire vocabulary in written input through noticing unknown words, requesting modified input by clicking on them, and receiving modified input. In other words, the CALL materials create opportunities for noticing vocabulary by glossing particular words and expressions that are expected to be troublesome for learners' comprehension of the text. This design has become commonplace in CALL materials such as those produced by Transparent Language and those that can be developed through authoring systems such as *LIBRA,* a CALL authoring tool based on a HyperCard application that can also be used to create controlled research experiments.

You, the reader, can participate in the activity by reading an example text as a learner would in English, French, German, or Spanish. As the learner, you can read the text, request definitions for (selected) unknown words, take the posttest and see your results. You can then look at your data and the analysis from the researcher's perspective.

The reading passages were all taken from online versions of newspapers and slightly adapted, that is, shortened to fit within the scope of this paper. Native or near-native speakers of each language identified the words they felt to be more difficult and crucial for understanding of the passage, created the glosses in consultation with a dictionary, and produced the reading comprehension and vocabulary questions.

The materials were authored through use of frames for the page layout, JavaScript for data gathering and analysis, and basic HTML coding for the content pages. The page layout was implemented through the use of three frames. The top frame serves as a container for data collection, the second frame displays instructions and glosses, and the third frame displays the reading passage and the post-test.

Since future access to the data was not required, data are collected through variables in JavaScript rather than submitted to a database (e.g., Microsoft Access or FileMaker Pro) via a CGI (common gateway interface) script. In other words, the data are stored on the client computer and deleted once the user exits the browser. Throughout the activity, JavaScript functions are used to keep track of the glosses that are invoked, to analyze the data, and to generate the student report and the analysis and report for the researcher.

Figure 2 shows part of the code in the <HEAD> of the reading passage. This example segment contains the glossed items (Target[1] — Target[4]) and the respective definitions (Htext[1] — Htext[4]). Target

and HText are JavaScript variables that are set up as an array. In addition to the actual item (e.g., currency), the HTML markup determining the display of each item is included (i.e., *Currency)* will be displayed in bold (<b>).

```
Target[1] = "<b>Currency</b>";
 HText[1] = "<b>Definition</b>: money, legal tender, e.g., US <b>Dollars</b> or Japanese <b>Yen</b> .";
 Target[2] = "<b>recalcitrant</b>";
 HText[2] = "<b>Definition</b>: formal adjective; refusing to obey or be controlled, even after being punished: recalcitrant behavior ";
 Target[3] = "<b>euro</b>";
 HText[3] = "<b>Definition</b>: the new currency in Europe ";
 Target[4] = "<b>launching</b>";
 HText[4] = "<b>Definition</b>: to begin (an activity, plan, way of life) ";
```

Figure 2. Example of HTML code showing glosses

Each glossed item is hyperlinked in such a way that when it is invoked (clicked), a JavaScript function is called that causes the definition of the word to appear in the frame above the reading passage.

When the learner clicks on "Please click here when you are ready for the questions," another JavaScript function is called which stores all data generated while the learner was going through the reading passage (i.e., words clicked on, frequency of clicks on particular words, and the actual items as well as the definitions) in the top frame. Additionally, in preparation for the correlation analysis, an array is set up in such a way that the value of the items clicked on is set to one (1) while the value of the items that were not selected is set to zero (0).

In this particular example, the word currency is the first Target, or Target[1]. As students read the passage and click on various words, the data are collected in the form of a comma-separated string of numbers, with each number representing a word. The data are stored sequentially, which means that every instance of a mouse click is recorded in order for later cross-referencing. For example, the string 1,1,2,4,1, indicates that the learner clicked on currency twice, then on recalcitrant, and then on launching before clicking on currency again. In preparation for the correlation analysis, the value of these items (1, 2, and 4) is set to 1. When the student takes the post-test, the responses to the vocabulary items are evaluated and stored as either correct (1) or incorrect (0). Both arrays of zeros and ones are then used to compute the correlation coefficient.

Since JavaScript was used to implement this example, and since JavaScript is interpreted differently by the two most popular browsers, careful programming which would not exclude one or the other was necessary. O' Reilly's *Definite Guide to JavaScript* (Flanagan, 1998, 3rd ed.) proved helpful in avoiding browser-dependent functions. While this was the most challenging feature to implement, other areas need to be cautiously considered when creating Web-based language learning activities. For example, materials developers must be aware the variety of displays learners will use to access materials. While monitors come in sizes ranging from 13" to 21", the more meaningful measurement is the monitor resolution, which indicates the total number of pixels on the screen. For example, a graphic which appears to be "the right size" on a high resolution monitor (e.g., 1280 x 1024) will appear much bigger on a standard 14" monitor running at 640 x 480. Authoring applications such as the cross-platform Macromedia Dreamweaver or Adobe's GoLive make it easy for developers to preview Web sites in various browsers and in various

resolutions. For a more detailed discussion of browser and display issues, refer to Niederst (1999) or Lynch and Horton (1999).

Implementation

Figure 3 outlines elements of the process of working with the CALL materials. The five columns represent the learner's action (Student), the data collected (Data), the implementation (Program), the type of inferences made (Inferences), and the researcher's observation (Researcher). The eight rows are arranged chronologically, representing learner's process while completing the activity.

	Student	Data	Program	Inferences	Researcher
1	Enters reading	Time	JavaScript		
2	Clicks on glossed items	Sequence & frequency	HTML (href) JavaScript	Student notices linguistic input	
3	Requests questions	Time	JavaScript	Student completes reading passage	
4	Responds to questions	Responses	JavaScript (answers are stored for subsequent analysis)	Degree of student comprehension	
5	Submits answers	Responses	JavaScript (evaluates responses	Student completes questions	
6			JavaScript (computes correlation)	Relationship b/w mouse clicks & correct answers	
7	Views report		JavaScript (generates and prints report)		Views report
8			Prints correlation		Observes correlation

Figure 3. Schematic overview of the process of using the CALL materials.

The first row illustrates how a learner, after reading the instructions, enters the reading passage. At this time, a JavaScript function records the time of entry and stores it the invisible top frame used to store all data collected.

Then, learners read the text and may or may not click on glossed items (row 2). If they do, two simultaneous actions occur: a) the requested gloss appears above the reading passage in the third frame and b) the mouse-click is recorded through another JavaScript function. Clicking on a glossed item is inferred to indicate that a learner has noticed an unfamiliar lexical item. The actions outlined in two may go through multiple iterations as the learner clicks on a number of vocabulary items and perhaps some more than once.

After completing the reading passage, the learner requests the vocabulary and comprehension questions (row 3). Again, the elapsed time is recorded as is the time when the learner begins with the questions. While elapsed time is not used for subsequent analysis in this particular example, it is sometimes used to

address other research questions (e.g., Jamieson & Chapelle, 1987). As the learner responds to the 14 vocabulary and 6 reading comprehension questions, the responses are stored for later analysis.

When the questions have been completed and the answers submitted (rows 4 and 5), the program stores the responses and evaluates them for correctness. The degree of student comprehension is inferred from the number of correct responses. The relationship between mouse clicks and correct answers can be inferred when looking at the correlation coefficient, which is computed based on the glosses that were selected and learner performance on the respective vocabulary comprehension questions (row 6).

The program then provides the learner and the researcher with the opportunity to view a report (Row 7). The learner can view a performance report along with a list of words for which input modification was requested, which can be used as a study guide. The researcher can view the same report and observe the correlation (Row 8). The information gathered might be used not only for the student and the researcher, but also for the teacher who wishes to tailor instruction to meet specific student needs.

The Empirical Research

The implementation is designed to investigate the following question: Is learners' noticing of vocabulary and receipt of modified input during the reading activity related to their retention of word meaning?

For the purpose of this demonstration, retention of word meaning is operationalized through performance on vocabulary items on a reading and vocabulary test that is part of the reading activity. In a research setting, one would also want to assess learners' knowledge of the target vocabulary before the activity, and to include a delayed post test to assess retention over time. What this example shows, however, is a methodology for assessing the extent to which noticing (with modified input) of particular words is related to their retention. Noticing is the independent variable which can have a value of 1 (word noticed) or 0 (word not noticed). Word retention is the dependent variable which can have the value of 1 (word retained) or 0 (word not retained). The unit of analysis, then, is the word rather than the learner, and for each of the target words, two values are obtained. For the complete set of target words, a phi correlation is calculated to summarize the relationship between the words noticed and those retained. The phi correlation is interpreted like other correlations (e.g., the Pearson product-moment correlation); that is, it can attain values anywhere within the range of -1.00 to 1.00 and should not necessarily be interpreted as indicating causality. The closer it is to 1.00 the stronger the relationship between two variables.

Because the unit of analysis is the word rather than the learner, a phi correlation is calculated for each learner rather than for the group of learners. To summarize the degree of relationship between noticing and word retention for a group of learners, the researcher can use the mean phi correlation for the group. In other words, a phi correlation is calculated for each learner, and then the mean phi is calculated by summing the individual correlations and dividing by the number of participants.

You can view the data you generated as you worked through the example and you can see how they are analyzed. Before choosing to view the data and data analysis, you should have worked through the reading activity from the learner's perspective. If you choose to view the data from the researcher's perspective before you have created the data as the researcher, some simulated data from the English example will be generated and displayed. Click here to see the data and analysis.

METHODOLOGICAL ISSUES

Methodological issues in this research design center around two related questions:

- To what extent does the on-line reading task create conditions for noticing linguistic problems and

receiving modified input?
- To what extent are the mouse click data valid measures of noticing?

Conditions for Noticing and Modified Input

Two related concerns should be considered about the validity of the conditions for noticing and receiving modified input. The first focuses on the way the task has been constructed to operationalize the constructs of noticing and modified input as they have been defined in interactionist theory. This is particularly important because the on-line task is different from the paper and pencil or the face-to-face tasks that have illustrated these constructs in other research. Moreover, noticing is typically discussed in relation to grammatical forms in the input rather than vocabulary even though the noticing hypothesis does not limit the hypothesized beneficial effects of noticing to morphosyntactic phenomena. Second, much of the research on noticing has manipulated the input externally, either through the materials or the teacher, but noticing is internal to the learner and therefore should be expected to equally effective whether it is motivated externally or by the learner. Modified input has typically been operationalized through oral face-to-face communication, but again here there is nothing about the construct of modified input that says it cannot occur in written texts.

Conclusions drawn about vocabulary retention in this context assume that learners noticed vocabulary that was causing difficulty for their reading, that is, vocabulary that they did not already know. A pretest would help to distinguish words that the learner might learn during the reading task from those already known. The former words are the ones of interest, and therefore they are the ones that should be included in the analysis when the correlation between noticing and correct responses is calculated. When the process of pretesting, observing, and postesting is done automatically, the computer can isolate the specific linguistic items that the learner needs to learn. In classroom research, this approach to assessing acquisition of noticed forms was used by Swain (1998), who constructed assessments specifically for individual students to assess their knowledge of the linguistic elements that researchers saw them focus on during task completion. Isolating the relevant linguistic forms for outcome assessment cannot be accomplished without observation of learners as they complete a task, and therefore it is ideally suited to CALL.

The second concern about conditions is the extent to which learners participate in the learning condition as it has been defined. Addressing this concern, Hulstijn (1997) suggests that a computer-assisted research task be followed by a retrospective interview or questionnaire to assess the extent to which learners participated, but the most reliable indication comes from the computer-documented data indicating the processes through which learners complete a task. DeKeyser (1995), for example, noted that records of learners' progress through the teaching materials as well as the retrospective data obtained from subjects' explanation of their strategies indicated that the intended implicit or explicit teaching methods he attempted to investigate were "sometimes overridden by their learning strategies" (DeKeyser, 1995, p. 398). In the case of noticing and receiving modified input, the effectiveness of the condition depends entirely on the learners' participation, and therefore the assessment of their noticing is essential to evaluating the quality of the learning condition.

Learners' participation depends on factors such as their ability level, their interest in the reading passage, and their desire to learn the language of the reading passage. These issues are inherent problems for constructing any good language materials; however, they are underscored when data from learners' use of materials is to be used to make inferences about their noticing. If learners have no desire to acquire the target language (as one may suspect is the case in laboratory settings) or if they have no interest in the meanings conveyed in the particular reading passage, they may not care enough to notice unknown vocabulary and to request modified input. Such a situation is disastrous for both learners, who waste their time, and researchers who fail to collect the relevant data.

Validity of Measurement

As Schmidt put it, noticing is an internal factor, which means that it is not observed directly, but must be inferred from observation of behavior. In the research described above an inference is made on the basis of observed mouse clicks--requests to see the modified input of the gloss. However, because this is a form of measurement, it is subject to the same concerns about validity as any other form of measurement (Chapelle, 1996). Validity of assessment of noticing should be considered in light of current principles of validation (Chapelle, 1999). These theoretical principles are put in the form of heuristics for evaluation through Bachman and Palmer's (1996) qualities of test usefulness, which suggests that a test should be evaluated against six qualities: reliability, construct validity, authenticity, interactiveness, impact, and practicality. In considering or developing a particular measure such as assessment of noticing through mouse clicks, the researcher needs to consider each of these qualities on the basis of a reasoned judgement.

Questions about *reliability* address the extent to which the assessment obtains a consistent picture of performance by minimizing irrelevant, or unmotivated variation across samples of performance. Factors such as instructions, and consistency of the input to the test taker, their response, and scoring procedure influence judgements about the potential reliability of a test. In this research design, because each assessment of noticing is obtained from item-level data which are correlated with a dependent variable (i.e., performance on each vocabulary item), any single assessment would not be considered reliable; however, the complete research design relies on multiple assessments of noticing, which together should be expected to be reliable. Since the item-level assessments of reliability are not added up to yield a score of how much of a "noticer" an individual is in the suggested design, the calculation of an internal consistency reliability across items is not relevant to this use of the measure.

C*onstruct validity* issues concern the clarity of the inference to be made, and the relevance of the assessment tasks and the scoring procedure to the inference. Construct validity requires that inferences be well defined, the knowledge of the participants be adequate so as not to bias performance, and the test tasks and scoring be consistent with the inference. The construct of noticing has been defined as directing attention to language. The behavior of requesting definitions during reading should indicate the readers' attention is being directed to the language, assuming the reader is seriously attempting to read the text.

Authenticity is evaluated by judgements about the extent to which the characteristics of the assessment reflect activities that the learner would engage in beyond the assessment setting. Use of hypertext during reading is authentic to some kinds of reading that the learners would be expected to engage in; however, the hypertext aspect of the reading would be different from many of the readings that learners would be doing. The journalistic style of the text would be authentic to some of the reading that learners would engage in, but not others.

Interactiveness is evaluated by assessing the extent to which the assessment engages participants' knowledge, communicative language strategies, and interest in the tasks. These features of interactiveness would be expected to be engaged during the reading activity, and could be prompted through the use of this text as one part of a unit that included other activities related to the European currency. A text such as this should not be used as a stand-alone, but as part of a task requiring learners to use the information they learn from the text.

Impact refers to the extent to which the assessment can be expected to positively influence students, language classes and programs, and society. Impact presents a different dimension for a validity argument than the other forms because it involves hypotheses directed beyond the assessment (Alderson & Wall, 1993; Bailey, 1996; Wall, 1997). Potential positive impacts on the learners is their learning some vocabulary, their learning to use hypertext for vocabulary learning, and their receipt of feedback on

retention after reading. Positive impact on teachers is that the materials allow for evaluation of their appropriateness by allowing them to observe the extent to which learners are noticing vocabulary as they read. A text with hypertext glosses can only be effective for vocabulary acquisition if it contains some words that the learners need to learn. Texts that do not record learners' performance provide no evidence to teachers about appropriateness. This assessment of noticing provides a means of empirically testing the hypothesis that noticing and receiving modified input in this format is valuable for vocabulary retention.

The materials also may have some negative impacts such as discomfort felt by some learners who are not accustomed to reading on a computer screen and those who may be uncomfortable having data gathered on their performance. Issues of privacy and consent involved in most empirical research on language teaching are particularly salient when SLA research is conducted through CALL tasks. Researchers need to decide the best way to obtain consent for data collection. The solution suggested here is to provide learners with a general notice about their responses being recorded to report to them before they enter the reading passage. This general notice, like those that appear in many Web applications, provides an implied consent for learners who chose to continue with the activity. At the end of the activity, learners could be asked if their data can be stored along with those of other readers for subsequent analysis without their name attached. At this point they can make an explicit choice about how their data are to be used. This approach to consent provides a solution which assumes the value of using the computer's data-gathering capabilities to help make the learner more aware of their learning and needs.

Practicality is assessed through consideration of the feasibility implementing the assessment given the available resources. Implementation of this assessment on the Web makes it practical for a wide audience provided technical issues such as browser and display compatibility are addressed. According to Which Browser, a browser statistic site, Netscape Navigator and Microsoft Internet Explorer currently account for 87% of browser use while AOL covers about 12% of the market, leaving less than 2% to other browsers. Clearly, such statistics are important for Web-based CALL materials developers since they need to know what works on which browser and design an activity accordingly. Implementation decisions must also be informed by the fact that browsers are implemented slightly differently on different platforms. For example, Niederst (1999) notes that dynamic HTML (DHTML) is problematic on Internet Explorer 4.0 on the Macintosh. This assessment was implemented using principles that would execute successfully on both browsers and across platforms.

This usefulness analysis is directed toward the method of assessing noticing as it is outlined in this paper. As the assessment is used in research and classrooms, additional evidence can be obtained to support these judgements about usefulness. For example, to provide additional support for construct validity, learners might be asked to think aloud while they are performing the reading task to see whether or not learners click on words that they don't know but need in order to understand the meaning of the text. Such data have been reported in studies of learner's use of hypertext. For example, Park's (1994) study investigating use of ESL multimedia through think aloud data identified thoughts such as "I think I have a lot of vocabulary that I don't know" (p. 147). This statement was made while the learner was clicking on one of the words in the input provides empirical support for construct reliability.

Table 1. Summary of Usefulness Analysis for Assessment of Noticing

Quality	Positive Attributes	Negative Attributes
Reliability	Multiple scores for noticing are used in the research design.	Noticing score for each word may be based on a single instance of behavior.
Construct Validity	Observed behavior appears to match construct definition of noticing language due to comprehension difficulties. Any serious reader who clicked on the words would be noticing them.	
Authenticity	The activities of reading in a hypertext environment and reading while looking up words are authentic to some of the reading contexts of language students.	The activity of reading with hypertext glosses is unlikely to be authentic to much of the target language reading that learners engage in.
Interactiveness	Communicative language strategies would be expected to be engaged for text comprehension.	
Impact	Learners may learn vocabulary and learn about using hypertext for acquiring vocabulary Learners receive information about their performance on the vocabulary and comprehension questions. Teachers can see to what extent learners are noticing and receiving modified input to evaluate the appropriateness of the materials. The profession can learn the extent to which noticing and receiving modified input correlates with immediate retention of vocabulary.	Learners not accustomed to reading on a computer screen may not be comfortable participating in this activity
Practicality	The entire process from instruction to data analysis is included in one program which can be run on any browser.	

CONCLUSION

Construction of computer-assisted reading materials points out that a "gloss" can take a variety of forms in multimedia materials. Roby (1999), for example, proposes that glosses can be characterized in terms of authorship (e.g., whether learners or materials developers wrote them), presentation (e.g., whether learners have access to them prior to or during the reading exercise), function (e.g., whether they include procedural or declarative information), focus (e.g., whether they make reference to the text or present new information), language (e.g., whether they were written in the target language or in the first language), and form (e.g., whether they provide verbal, visual, or audio help). Each of these provides learners with a variety of input modification and raises the empirical question of whether some types of modifications are more effective than others for learners. This question must be addressed through the type of computer-

assisted methods described in this paper because the question is not only what kind of glosses learners have access to, but also which glosses learners actually use and whether their use correlates with vocabulary retention.

The study of on-line noticing of vocabulary in CALL materials is just one way that hypotheses from SLA theory can be tested through CALL design with built-in data collection and analysis. This paper has explored the methodological issues associated with such research to sharpen understanding of the technical and measurement issues associated with linking SLA and CALL. Future work in needed to apply these methods in research and instruction in order to produce results that contribute to SLA theory and CALL practice.

ACKNOWLEDGEMENTS

We are grateful to the GRApES research group for comments on an earlier version of this paper, to the four anonymous reviewers for their insightful comments and suggestions, and to the following people for their help with various parts of the examples: Arati Bhat Manjeshwar, Bellinda Hegelheimer, Julio Rodriguez, Cindy Myers, and Douglas Mills.

The CALL materials created for this paper draw on JavaScript examples by Douglas Mills and Marmo Soemarmo.

ABOUT THE AUTHORS

Volker Hegelheimer is an assistant professor in the Department of English and the Program in Linguistics at Iowa State University. He teaches ESL and graduate courses on technology in language teaching and research and his research interests include applications of the WWW and emerging technologies in language learning and language testing.

E-mail: volkerh@iastate.edu

Carol A. Chapelle (Ph.D. University of Illinois at Urbana-Champaign) is Professor of TESL/applied linguistics at Iowa State University. She teaches courses in applied linguistics and ESL and conducts research on computer-assisted language learning and language testing. Her book, *Computer Applications in Second Language Acquisition: Foundations for teaching, testing, and research* (Cambridge University Press, forthcoming), reviews past research on computer applications and charts directions for second language acquisition in the era of technology. Her papers have appeared in journals such as *TESOL Quarterly*, *Language Learning*, *Language Testing*, *and Language Learning & Technology*. She is editor of *TESOL Quarterly*.

E-mail: carolc@iastate.edu

REFERENCES

Alder son, J.C., & Wall, D. (1993). Does washback exist? *Applied Linguistics 14*, 115-29.

Bachman, L.F., & Palmer, A.S. (1996). Language testing in practice: Designing and developing useful language tests. New York: Oxford

Bailey, K. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, *13*(3), 257-279.

Chapelle, C. A. (1996). Validity issues in computer-assisted strategy assessment. *Applied Language Learning*, *7*(1), 47-60.

Chapelle, C. A. (1998). Multimedia CALL: Lessons to be learned from research on instructed SLA. *Language Learning and Technology, 2*(1), 22-34. Retrieved May 12, 2000 from the World Wide Web: http://llt.msu.edu/vol2num1/article1/index.html.

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal, 80,* 183-198.

DeKeyser, R. M. (1995). Learning second language grammar rules: An experiment with a miniture linguistic system. *Studies in Second Language Acquisition, 17,* 379-410.

Desmarais, L., Duquette, L., Renié, D., & Laurier, M. (1998). Evaluating learning interactions in a multimedia environment. *Computers and the Humanities, 22,* 1-23.

Doughty, C. (1987). Relating second-language acquisition theory to CALL research and application. In W.F. Smith, (ed.), Modern Media in Foreign Language Education: Theory and Implementation, (pp. 133-167). Lincolnwood, IL: National Textbook Company.

Doughty, C. (1991). Second language instruction does make a difference: Evidence from an empirical study of SL relativization. *Studies in Second Language Acquisition, 13,* 431-469.

Ellis, R. (1995). Modified oral input and the acquisition of word meanings. *Applied Linguistics, 16*(4), 409-441.

Flanagan, D. (1998). *JavaSCRIPT: The definite guide (3rd ed.)*. Cambridge, Massachusetts: O'Reilly.

Gass, S. (1997). *Input, interaction, and the second language learner.* Mahwah, NJ: Lawrence Erlbaum Associates.

Gass, S. M., & Madden, C. G. (Eds.). (1985) *Input in second language acquisition.* Rowley, MA: Newbury House Publishers.

Hegelheimer, V. (1998). Effects of textual glosses and sentence-level audio glosses on online reading comprehension and vocabulary recall. Unpublished doctoral dissertation. Department of Educational Psychology, College of Education, University of Illinois, Urbana, IL.

Hsu, J. (1994). *Computer assisted language learning (CALL): The effect of ESL students' use of interactional modifications on listening comprehension.* Unpublished doctoral dissertation, Department of Curriculum and Instruction, College of Education, Iowa State University, Ames, IA.

Hulstijn, J. (1993). When do foreign language learners look up the meaning of unfamiliar words? The influence of task and learner variables. *Modern Language Journal, 77*(2), 139-147.

Hulstijn, J. (1997). Second language acquisition research in the laboratory: Possibilities and limitations. *Studies in Second Language Acquisition, 19,* 131-143.

Jamieson, J., & Chapelle, C. (1987). Working styles on computers as evidence of second language learning strategies. *Language Learning, 37,* 523-544.

Jourdenais, R., Ota, M., Stauffer, S., Boyson, B., & Doughty, C. (1995). Does textual enhancement promote noticing? A think-aloud protocol analysis. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (Technical Report #9) (pp. 183-216). Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Center.

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

Long, M. H. (1988). Instructed interlanguage development. In L. Beebe (Ed.), *Issues in second language acquisition: Multiple perspectives* (pp. 115-141).New York: Newbury House.

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

Long, M. H., & Robinson, P. (1998). Focus on form: Theory, research and practice. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 15-41). Cambridge: Cambridge University Press.

Lyman-Hager, M., Davis, J., Burnett, J., & Chennault, R. (1993). Une vie de boy: Interactive reading in French. In F.L. Borchardt & E.M.T. Johnson (Eds.), *Proceedings of the CALICO 1993 Annual Symposium on Assessment* (pp. 93-97). Durham, NC: Duke University.

Lynch, P., & Horton, S. (1999). *Web style guide: Basic design principles for creating web sites.* New Haven: Yale University. [Available on the World Wide Web at http://info.med.yale.edu/caim/manual/]

Martinez-Lage, A. (1997). Hypermedia technology for teaching reading. In M. Bush & R. Terry (Eds.), *Technology enhanced language learning* (pp. 121-163). Lincolnwood, IL: National Textbook Company.

Niederst, J. (1999). *Web Design in a Nutshell.* Cambridge, MA: O'Reilly.

Park, Y. (1994). *Incorporating interactive multimedia in an ESL classroom environment: Learners' interactions and learning strategies.* Unpublished doctoral dissertation, Department of Curriculum and Instruction, College of Education, Iowa State University, Ames, IA.

Pica, T. (1994). Research on negotiation: What does it reveal about second-language learning conditions, processes, and outcomes? *Language Learning, 44*(3), 493-527.

Robinson, P. (1995). Review article: Attention, memory and the "noticing" hypothesis. *Language Learning, 45*, 285-331.

Roby, W. (1999). What's in a gloss? *Language Learning & Technology. 2* (2) 94-101. Retrieved May 12, 2000 from the World Wide Web: http://llt.msu.edu/vol2num2/roby/index.html.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129-158.

Schmidt, R. (1992). Awareness and second language acquisition. *Annual Review of Applied Linguistics, 13*, 206-226.

Schmidt, R., & Frota, S. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. Day (Ed.*), Talking to learn: Conversation in second language acquisition,* (pp. 237-326). Rowley, MA: Newbury House.

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

Swain, M. (1998). Focus on form through conscious reflection. In C. Doughty & J. Williams (Eds.), *Focus on form*, (pp. 64-81). Cambridge: Cambridge University Press.

Wall, D. (1997). Impact and washback in language testing. In C. Clapham & D. Corson, (Eds.), *Encyclopedia of language and education. Volume 7: Language testing and assessment,* (pp. 291-302). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Watanabe, Y. (1997). Input, intake, and retention: Effects of increased processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition, 19*(3), 287-307.

White, J. (1998). Getting the learners' attention: A typographical input enhancement study. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 85-113). Cambridge: Cambridge University Press.