ReCALL **14** (1): 146–166. © 2002 Cambridge University Press
DOI: 10.1017/S0958344002001210 Printed in the United Kingdom

CALL and the classroom: the case for comparative research

PAUL ALLUM
Rikkyo University, Nishi-Ikebukuro 3-34-1, Toshima-ku, Tokyo 171-8501, Japan
(*email: pallum@gol.com*)

Abstract

Educators are more and more concerned to integrate human and computer capabilities as efficiently as possible. There is an increasing demand in language teaching for evidence on which to base decisions about use of this ever more prevalent tool. One main yardstick for evaluation of CALL is, naturally, the human teacher. Thus there is demand for comparative data. Attempts to provide such data in a way that allows useful generalisations to be made or accurate analyses of the outcomes are still comparatively limited and difficult to produce. Many attempts have been strongly criticised and the very concept of comparative studies of this kind has been attacked. The belief asserted here is that there is still a need for general studies in a variety of contexts. A longitudinal comparative study is reported that tries to take into account some of the major criticisms, while providing support for the idea that these kinds of study still need to continue. The study provides a detailed description of the design and implementation, an evaluation in terms of Second Language Acquisition (SLA) criteria, and pre and post tests to measure outcomes across a range of language learning tasks.

1 Comparative studies – Computer Aided Instruction (CAI) versus 'conventional classroom'

1.1 The aims of comparative research

Early CAI projects needed to justify the expenses involved in their development and were therefore keen to show that better results were obtained compared with the cheaper alternative of conventional classroom delivery (Levy, 1997:13–28). For example, such a need is clearly expressed in Schrupp *et al*. (Schrupp, Bush and Mueller, 1983:17) in their evaluation of CAI-based interactive video compared with simple video. "The purpose of this experiment was to measure any quantifiable advantage in learning outcome attributable to computer-assisted interaction."… "It is very important to justify the considerable costs involved by showing that this technology can improve learning outcome."

The current situation is very different. Networked computers are widely diffused, especially in higher educational settings. In many circumstances, the introduction of

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

CALL would involve little or no hardware costs, and relatively minor software costs. This should have changed the focus of much comparative research. There is a need to show whether and how CALL is just as effective at delivering instruction as conventional means. Such delivery is likely to be cheaper than employment of teachers to do the same work. Equally, there is a need to clarify if and where there are curriculum areas in which CALL cannot perform as well as conventional means. This approach is clearly suggested by Clark (1994b:22) in his discussion of comparison of various media as means of instructional delivery.

> "*We need to ask whether there are other media or another set of media attributes that would yield similar learning gains*. The question is critical because if different media or attributes yield similar learning gains and facilitate achievement of necessary performance criteria, then in a design science or an instructional technology, we must always choose the *less expensive way*" (my italics)

Despite this, Chapelle (2001:26) quotes Richard Tucker, in his 1999 address at the CALICO symposium as follows.

> "We must implement as soon as practical a multifaceted and longitudinal research agenda to examine the value added to students who pursue some or all of their language education using innovative technologies."

She adds "The issue is how to define 'value added' and how to assess the extent to which value has been added for learners who use CALL."

This seems to follow the same line of thinking as that expressed in the evaluation of early research, that is, comparative research must show that CALL adds value over and above that offered by conventional means. The current research rejects this approach for the reasons given in the second paragraph. While admitting that it would be useful and important for educators to know that CALL can perform better in some kinds of instructional delivery, the main aim here is to confirm that CALL does indeed deliver as effectively as conventional means in a range of language learning tasks. More importantly, in the longer term, the aim is to discriminate where such is and where such is not the case. This should allow principled integration of CALL and classroom in the most effective way, apportioning those parts of the curriculum to CALL that can be more economically delivered through this means. Such a goal is aimed at adding value not by proving that CALL is better, but by using always limited and expensive teacher time more effectively. A possible way to implement this is giving preparatory, controlled practice or review work to CALL, while using class time for the freer production of comprehensible output and negotiation of meaning that SLA interactionist theories consider vital for SLA, an area in which CALL still lacks sufficient flexibility or power.

1.2 The current situation

Levy (2001: 6) reports that comparative studies have recently been an increasingly popular type of research in CALL – 17% of all CALL-related publications in 1999, 30% in 2000. He indicates that for this two-year period "the most common kind of comparison

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

is the one where the experimental group uses the computer or technology in some way, and the control group has 'regular classes'." This indicates that teachers and researchers still feel the need to justify decisions to use CALL despite quite a range of research that shows CAI or CALL are at least as effective as 'conventional teaching' in some ways. In a review of comparative studies Clark (1985c) indicates that when methodology is kept consistent, there is no difference in results between computer-based instruction and teacher-led instruction. Pederson (1988:104) gives several examples of comparative CALL projects that have showed no difference. The desire to compare is not even dampened by the considerable criticism that such comparative work has attracted.

1.3 Criticisms of comparative research

Chapelle (2001:44) agrees with Clark (1994b) that some researchers may be motivated by an approach that takes an apparently desirable solution – CALL – and then tries to find problems in instructed SLA that can be solved by it. Such motivation may result in "gross comparisons between computer-using learners with those learning through other media (e.g. Adair-Hauck *et al.,* 2000; Nutta, 1998), an approach unlikely to shed light on the problem or the solution." Pederson (1988:106–7) has argued that such studies have little value because, amongst other reasons, they cannot be accurately replicated and close linkage of independent and dependent variables is difficult if not impossible. Thus results cannot be generalised. Levy (2001:7–8) criticises such studies from another point of view, lack of clarity in the definition, description or control of relevant variables by the researchers. He mentions such considerations as whether the teacher was present in the computer laboratory and what role he/she played, novelty effects in short-term cross-sectional studies, lack of information about the CALL program, and lack of information about the content of the activities of the control group in the 'traditional' classroom. Pederson (*ibid*:105–5) makes the same point and refers specifically to several studies. Levy further quotes Murray (1999:80) as saying that such comparisons provide little information that is useful to guide future development and use, a view supported by Pederson (*ibid*.), and are based on the false assumption that the media can be varied without changing instructional content or strategy. Clark (1994b) argues that the medium itself is irrelevant. It is the method and content that determines what is learnt. He claims that any "necessary teaching method could be designed into a variety of media presentations." (ibid.:22)

1.4 Means, media and methodologies

Pederson (1988:100) claims "Recent CALL research, however, suggests that the design of computer software to cause adjustments in cognitive processing, not the *medium* used to deliver instruction, stands the best chance of affecting learning outcomes." She refers to her own research (Pederson: 1986). In a French reading exercise delivered through the computer, she compared the effect of passage availability versus non-availability during the answering of comprehension questions on retention of contents of that passage. The results showed that non-availability improved retention of the content. This result tells us nothing specific about CALL. The same experiment could have been performed with paper and pencil. Pederson's reason for using CALL was its ability to reliably hide the text when students answered the questions, something that apparently

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

could not be ensured with students in class. Thus CALL is the medium, but the method and results are not CALL-specific. This research provides a practical reason for using CALL in larger classes where the teacher cannot monitor the students. As Pederson says, CALL *facilitates* certain kinds of learning. A similar comment might be made on Chapelle and Jamieson's (1986) experiment. In a comparison of the different usage and learning outcomes for various types of students studying ESL through PLATO, she found that there was a strong negative correlation between field-independence and both time spent on and attitude toward studying with the system. Field-independence was, however, strongly correlated with improvement in TOEFL score and an oral test. In explaining the negative relationship she says as follows:

> "It is likely that the FI students, who are capable of and accustomed to using their own internal referents, found the structured approach of the lessons in the ESL PLATO series to be inconsistent with their learning styles. They may have found it irritating to have information and exercises structured in a way different from how they would have done it for themselves."

As with the Pederson research, there is little that can be said about CALL *per se*. The negative reaction to PLATO is ascribed to instructional design that is not CALL specific. Quite likely these students would have objected as strongly to having the same work foisted on them in the classroom.

Taking Clark, Pederson and Chapelle's views to an extreme might lead to the conclusion that there is almost no *specifically* CALL-related agenda in taught SLA, just as there is no particular text-book, video or language laboratory related agenda *per se*. For example, CALLware may be judged as more or less effective as it has the potential for and actually results in activities such as meaningful exchange in which there is a focus on form. A specific evaluation may involve use of specific CALLware but the results are indicative of qualities of instructional design, or the relationship between instructional design and learner attributes, as representative of underlying cognitive variables, not of CALL-specific features or capabilities. In other words, we can say that this piece of CALLware lacks certain attributes and thus fails to cause certain cognitive processes, and that's the likely reason it fails, but such attributes may well not be specific to the medium. Similarly, under this approach, much CALL-modulated research can contribute to taught SLA theories, but has nothing specific to say about CALL. Pederson (1988:115–122) quotes research in which the computer acts as a useful means of obtaining data that would otherwise be hard to obtain, or is the means of instructional delivery, but where the results tell us nothing specific about CALL, except perhaps incidentally, though importantly, that it has the capability to deliver sound pedagogy in certain areas. In brief, no specifically computer-related variable is manipulated in the experiments quoted, which, in view of Clark's claim about media per se, might not be surprising. CALL becomes a means of research on taught SLA, effectively evaluating an instructional design by testing it through CALLware. This use of CALL is also strongly advocated by Doughty (1987).

1.5 Ways forward

The use of CALL as a research means and the evaluation of CALL on the basis of its

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

instructional design seem well justified. However, these approaches do not directly answer the question posed by this research – should CALL be used and, if so, how should it be used in a particular curriculum to give the best results? There are suggestions on ways forward.

> "The most convincing way to demonstrate the language learning potential of a CALL activity is through the study of learning outcomes. In other words, if learners were to have acquired particular grammatical forms or vocabulary through CALL task, then results of an assessment after learners have completed the task can provide some evidence of the language learning potential of the task. The evidence is much stronger, of course, if pre-test data indicate that the learners did not know the target forms before beginning to work with CALL. *Still stronger evidence is obtained if a contrasting group that did not use the CALL task or used the CALL task in another form failed to make similar gains*." (Chapelle 2001:74) (my italics)

Other ways forward in CALL evaluation have been suggested, though such suggestions may not always be specifically aimed at comparative studies. Dunkel (1991: 25–26) suggests that "Researchers interested in measuring *gains in achievement as a function of using computers to teach content or skill* learning should focus their research on examinations of specific conditions under which the use of computer materials results in L2 learning." (my italics) She suggests these four questions as worth pursuing:

- whether CALL is more effective for certain L2 skill areas than others;
- whether certain levels of proficiency profit more from computer use than others;
- what role feedback plays in the effectiveness of CALL programs;
- what degree of learner control is related to effective CALL designs.

Chapelle (2001:44–93) has strongly argued for evaluation that measures both the *potential* of a piece of software to provide conditions that engender language activities that have been suggested as important by SLA theory as well as the *actual* activities engendered in practice. She defines the former evaluation as 'judgmental'. This is contrasted with 'empirical' evaluation in which collected data indicates whether the potential identified in the judgmental evaluation is actually fulfilled in practice. She identifies six criteria derived from SLA research against which CALLware should be measured in either type of evaluation; language learning potential, learner fit, meaning focus, authenticity, positive impact and practicality. (Chapelle 2001:59, 68) For example, in a judgmental evaluation of how well the CALLware meets the criterion of language learning potential, the question to ask would be "Do task conditions present sufficient opportunity for beneficial focus on form?" In an empirical evaluation related to the same criterion, the question would be "What evidence suggests that the learner has acquired the target forms that were focussed on during the CALL task?" or "What evidence indicates that learners focused on form during the CALL task?"

The need to focus on specific SLA-designated variables in evaluation of CALL-delivered materials has also been strongly supported by Pederson (1988:126).

> In short, the research clearly indicates that CALL is highly context-bound. The

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

> nature of the learners, the task they believe they are supposed to perform, and the way the CALL materials are designed all play individual and interdependent roles in determining learning outcomes. Researchers should continue to investigate all three variables in a variety of contexts in future CALL basic studies.

Pederson *(ibid*.) suggests that, while we should forever abandon attempts to prove the superiority of computers over some other medium, there is still a role for comparative research.

> "The question must be, 'How do computers make a difference?' Appropriate CALL research is needed to provide *theoretical*, *explanatory,* and *practical* information about the specific ways the computer can deliver instruction that is particularly well-suited to specific language learners who, in turn, perform clearly defined tasks." (Pederson, 1988:108)

Clark (1994b:22) suggests a different way forward. Assuming the truth of his claim that any instructional medium can be replaced by another, he says our focus should change from"media attributes as causal in learning to media attributes as causal in the cost-effectiveness of learning."

Two main research thrusts can be identified, one that attempts to relate CALL and taught SLA theory more closely by using CALL to evaluate instructional designs based on SLA theory, and one that looks for the effectiveness of CALL, though they are by no means mutually exclusive. Among the latter type of research, there are suggestions that we should look for efficiency either by looking at how well the instructional design matches with designs predicated as effective by instructed SLA theory, or at whether greater speed and less cost result from use of CALL. Ideally, perhaps, we would want to evaluate CALL-ware on all counts: does it have a sound pedagogic design on the criteria designated as important by theory, does it deliver the learning activity in reality, and does it do so as effectively or more so than other means? Finally, we might want to ask as Pederson (1986) did, does CALL provide in practice a better or equivalent means of delivering a particular instructional methodology? 'Better' or 'equivalent' might be determined in terms of reliability, speed, practicality or cost in relation to outcome.

2 The current investigation – research design

The current research begins by attempting to design a comparative experiment that avoids some of the criticisms of lack of clarity made by Levy, fulfils some of the suggestions made by Chapelle in that it compares outcomes of CALL delivery with teacher delivery using pre and post tests, and relates tasks as far as possible to SLA theory. It also takes the first step towards answering the last question in the previous paragraph.

2.1 Purpose

The aim was to compare two means of delivery of the same material, using the same

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

fundamental methodology. The means were teacher, textbook and classroom versus delivery through computerised exercises on the Web. The long-term goal is to see how classroom work and outside Web-delivered homework should best be integrated. One main supposition is that much of the very limited classroom time is 'wasted' in the sense that the presence of a teacher is not especially important for much of the time during exercises. This is particularly true with 'preparatory' work – the introducing of new vocabulary or phrases – with many listening exercises, with some of the basic grammatical explanation and manipulation, and with review or check exercises at the end of a unit. In the detailed comparison of teacher-fronted and CALL delivery of one unit given in Appendix 1, I will point out these parts in more detail. At the same time it was hoped that the design would answer some of the criticisms alluded to in Section 1.4, and thus encourage the debate on comparative studies to develop further.

2.2 Period

The experiment ran over eleven weeks with once-a-week ninety-minute classes, approximately one 13 week semester, including a couple of days off for national holidays.

2.3 Software

All exercises for the CALL class were authored on Hot Potatoes 5.2. Standard functions of this package were used. The exercises were viewed in the Japanese version of Internet Explorer 5.0.

2.4 Subjects

There were 28 students in the CALL group, 26 in the classroom group. They were all first year university students, non-English majors, taking compulsory English. They had had six years of secondary school English education. Despite this, their level would be classified as false beginner or elementary in speaking, writing or listening ability. Simply put, their communicative ability was very limited. They had intermediate reading skills. Their six years would have introduced all the basic grammatical structures of English. Few had had a native speaker teacher, and fewer had been outside Japan or had any substantial contact with foreigners.

2.5 Course aims

The aim of the course was mainly to improve students' comprehension of the language they had learnt at school, while reviewing it in a more communicative context.

2.6 Conditions

The CALL delivery took place in the regular ninety-minute class time. All students were present together. There was a moderate amount of 'spin-off' talk. The teacher was present throughout in a virtually non-participatory role, except for the first two classes in

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

which considerable help was given to familiarise students with the software, but none with English.

2.7 Measures

Identical paper and pencil pre and post tests were given on the contents of each unit, and an identical pre and post test on the contents of all units was given at the very beginning and the very end. All measures were repeated. Measures, both pre and post, were taken at the beginning of the first of the two ninety-minute periods allotted to a unit. Thus there was a week between finishing study and post-test. Measures were taken in the following four main categories: vocabulary learning, listening comprehension, grammatical construction, functional phrases; and one lesser category: spelling. 'Main' indicates that the category was sampled most frequently and represented a core aspect of the curriculum. Not all measures were taken each time. More were gradually taken as the class became accustomed to the system and took less time to take the tests. A sample test is appended (Appendix 2). Additional data was obtained through a questionnaire on student attitudes to the material and method as well as on their sense of progress. Finally, there was a lot of informal observation of student behaviour in the CALL class.

2.8 Analysis

Analysis was confined to T-tests on the results of pre and post tests and comparative differences in improvement. For space reasons, only the latter are reported in detail in this paper.

2.9 Variable control

Assignment to class is pseudo-random: students are alternately assigned to each class, i.e. if Abe and Ando appeared in order on the list of first year students, one would go to class A, the other to class B. Marking of quizzes was done by a third party not involved in the project, with checks on about 10% carried out by the writer. The material was authored or taught by the writer. The materials were tightly controlled. Students did almost exactly the same exercises on the web page as they did in the book and heard the same accompanying audio CD. Graphics were exactly the same. Methodology was virtually the same. Simply put, the textbook was effectively digitised.[1] Further, time use was controlled. Students progressed through the syllabus at the same pace, two ninety-minute classes per unit, although there were variations on time per exercise. A comparison of the exercises of one unit is given in Appendix 1.

There are some potentially important points of difference. The first is in feedback - both the timing and quality. Feedback in the CALL exercises could be accessed at any time, for example after completing just one out of six questions or after filling in only three of eight blanks. Two main kinds of feedback were used in the Hot Potatoes applications. String checking in JQuiz (a free input answer application) answers informs the user what parts of an answer are correct, in other words, what words or part-words in the

[1]Permission for the authoring and delivery of these materials was obtained from the publisher.

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

answer match the predicted answer. In JCloze (a gap-fill application), JMatch (a drag and drop matching exercise) and JBC (a multiple-choice exercise) the user only discovers if the answer is right or wrong. All the exercises allow 'Clues' written by the author. Feedback using error anticipation can be produced but was not used, nor was the 'Hint' function that displays the letters of the answer one by one. As delayed production of the correct answer, for example after three tries, is not an option, the correct answer was not made available. In consideration of this, the questions were designed to be not too difficult. Thus most students kept correcting their errors until they got the right answer. In the classroom, feedback was generally after completion of each exercise, and correction only addressed specifically to those who called out answers. Such feedback, however, was often elicitatory, explanatory or requesting clarification, in contrast to the CALL feedback.

While the CALL group could play the audio as much as they wanted, and repeat whichever parts they wished, the class were in lockstep mode. While the number of playings by the teacher was based on experience and asking the students if they wished to hear it again, it is not clear whether all students had heard the CD as much as they wished. Thus audio input was closely controlled by the CALL group but not by the class group.

The CALL group had more control over the pace of their learning, though they had to keep in mind the time limit per unit.

The class group had more opportunities to speak out. Production in the CALL class was either monitored by the computer or handed in as written assignment. Thus spoken production with attendant negotiation of meaning with peers was a feature of the class but not the CALL group.

2.10 A comparative 'judgmental' evaluation based on SLA criteria

Chapelle (1997, 1998) has suggested SLA criteria for evaluating the potential of tasks. These are based on interactionist theories of learning. An attempt is made to apply these to the methodology as realised through these two forms of delivery. Reference will be made to the one unit shown in Appendix 1, which had the objective of teaching asking and giving of directions. All units followed a similar pattern, though some had more focus on grammar e.g. adverbial expressions of frequency, or regular/irregular past tenses. The stated overall objective was teaching 'asking for, giving and following directions.' Objectives by category were: functions – expressions used to ask for and give directions; grammar – prepositions of location, imperative mood; vocabulary – stores and services. Interested readers are referred to Chapelle (1997:27–28) for the basis of the following evaluation.

Making key linguistic features salient

Learning targets were salient, in terms of vocabulary and functional expressions. Syntax was not, perhaps, as salient. In tasks 1–7 syntax was not highlighted, though the matching in T1a required matching of prepositional phrase with picture, and these phrases were translated in T1b. In task 8, the table of typical directions, made available as a reference while doing the task, divided up the elements of the sentence into verb, object, preposition and prepositional phrase, or verb, complement. Thus the various prepositions were made clearer (to/second street, across/the river). But there was no highlight, and the sepa-

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

ration was not systematic (It's/on the corner). In the CALL class, the cloze exercise that replaced the information gap exercise focussed on both syntax and content (for example several prepositions were 'gapped') and thus drew attention to this specifically syntactic element. Although the imperative was used for all directions, there was no highlighting of this feature.

Offering modifications of linguistic input

There was modification of input in terms of visual cues (pictures) for both vocabulary and phrasal items, and some by definitions of vocabulary for both groups. Visual gestural cues from the teacher were also available to the class group. These were not really at student request. Repetition was likewise available for both, in theory by request for the class group (Japanese students rarely make such requests), and definitely in practice for the CALL group. The latter also sometimes had access to the 'Clue' function that modified the input. Additionally, translation may be seen as a modification in that it allows comprehension of the input. Both classes had this modification.

Providing opportunities for comprehensible output

There were clearly more opportunities for the class group to produce spoken comprehensible output (T6b, T7, T8). Output for the CALL group was limited to written form. However, the amount of such output may not have been so different except for T8. And it is hard to judge whether repeated input to get the correct answer by the CALL group falls into this category of opportunity.

Providing opportunities for learners to notice their errors

Pairwork and written work allow considerable opportunity to notice errors. In so far as the pace of written work may allow more time to notice, the CALL group may have had more opportunity.

Providing opportunities for learners to correct their linguistic input

While opportunities to correct exist for both, the CALL class have the advantage of constant attention being drawn to errors. The nature of the written input and evaluation strongly encourages correction. They are also free from the pressures to avoid correction in many communicative situations in which maintenance of discourse is critical.

Supporting modified interaction

A definition of what constitutes interaction that is useful for SLA is given by Chapelle (1998:27). "These interactions need to move the learner towards a task goal and stop progress along the way to focus attention on language." They can consist of mouse clicks or hypertext links in the case of CALL. This seems at first to mean that the constant efforts to type in the correct answer for the CALL group are also types of modified interaction. However, such modifications do not really occur on the way to achieving a task goal for the CALL students as they do for the class group. For example, in T8 the class group had to communicate the location of places on the map, but the CALL group only had to fill in the gaps in the cloze exercise. The class group's pairwork in T8 or T6b should have encouraged modification.

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

Two-way interaction with real communicative goals

As mentioned above, the CALL group had fewer opportunities to engage in two-way tasks for real communicative purposes.

The main differences could perhaps be summarised as follows. The CALL group encountered slightly more salient linguistic features. They had far more opportunity for modification of input in terms of repetition. They had fewer opportunities to produce comprehensible output in one sense, as mentioned above, and none of this is in spoken form. An exact assessment would entail recording how much was produced in the spoken exercises by the class group. They had equal opportunities to notice errors (though this too depends on how rich an environment pairwork provides for this compared with written work) and far greater opportunities to correct their linguistic input, though only in written mode, and not as a result of negotiating meaning to repair a breakdown. Modified interaction drawing attention to form *on the way to achieving a task*, was minimal for the CALL group, and they had little opportunity for two-way interaction in pursuing real communicative goals.

3 Outcomes

3.1 The pre and post test data

The class group had more opportunity to produce phrases in T6 and T8, scored significantly better in the dialogue dictation, and in the written dialogue (based on someone asking for and giving directions), not to mention a better (though not significantly so) score on the listening comprehension of a directions dialogue. Their written production of a directions dialogue was also significantly better. Yet in a simple translation test of the phrases they only scored equal to the CALL group and they showed significantly less accuracy on a grammar test (translation) that focussed on the prepositional phrases dealing with location (see Table 1). It is tempting to put the improved listening and written dialogue skills down to the greater communicative practice available to the class group, and the greater grammatical accuracy down to the focus on form likely to have occurred in the cloze exercise. The first of these would tie in well with SLA theory on the

Table 1. *Scores on Pre and Post Tests for Unit 5 (Dic = Dictation; Dia = Dialogue; Co = Comprehension; ++ = non-significant improvement at p<.05)*

	Class 1			Class 2 (CALL group)		
Measure	Before (%)	After (%)	Difference (%)	Before (%)	After (%)	Difference (%)
Phrases	58	77	19	65	86	21
Vocab	51	81	30	50	84	34
Grammar	29	36	8++	30	61	31
Spelling	56	68	12	58	70	12
Listen Dic/Dia	29	47	18	43	48	5++
Listen Co/Dia	21	48	26	23	40	17
Write Dia	16	41	25	22	33	11
Total Average	38	57	19	42	61	19

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

importance of real communicative tasks such as information gap exercises in acquiring new forms. Spada (1987, in Lightbown and Spada, 1999:144) shows that teaching grammar specifically, even though it may not be part of a communicative task, can improve measure of grammatical improvement relative to classes that do not have specific grammar teaching. This could explain this result. In Figure 1, improvement is expressed as a percentage. However, looking at the bigger picture of results, it soon becomes clear that one-time treatments need to be treated with great caution, the point made by Levy (2001) in his criticisms of CALL comparative research.

When results are viewed over a period, it becomes clear that the criteria related to SLA theory, if used to account for the data from Unit 5, cannot so easily explain the overall similarity in the results. Although the class group had consistently more opportunity to employ the language in a communicative situation, there was no consistent effect. This could be for two reasons: the measures were not suitable to reflect any increased skill they had; pair work does not lead to a high rate of modification of output or input – a definite possibility in large monolingual classes. The similarity in progress becomes even clearer when improvement is viewed per measure as in Tables 2 and Table 3.

In the comparison shown in Table 2, different measures of listening, dictation or comprehension, were collated into one figure.

The results overall confirm that the medium, if the two 'media' are teacher and CALL, is not especially important, at least for the measures taken and with students at this level. They also confirm that this is true across a range of exercises with different aims – teaching functional phrases, vocabulary, grammar, listening, and even spelling. Further, this is true even when simple CALL exercises are used, and the full power of CALL to provide more sophisticated feedback, branching, or higher quality visual information is not utilised. In addition, they show that this is true with the most rapidly produced

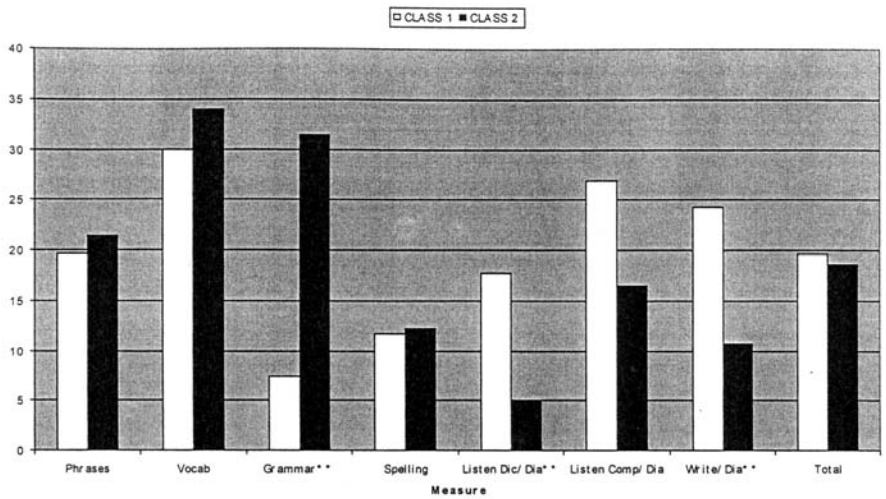


Fig. 1. Difference in improvemnt on Unit 5. (Two asterisks indicate a significant difference between the classes, p <.05. Class 2 is the CALL group. Dic = 'dictation'; Dia = 'dialogue'; Comp = 'comprehension'.)

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

Table 2. *Comparative improvement by measure expressed as percentage (** = significant difference)*

Measure		Unit 1 (%)	Unit 2 (%)	Unit 3 (%)	Unit 4 (%)	Unit 5 (%)	Total (%)
Phrases	Class 1	18	26	8	-1	19	14
	Class 2	22	31	6	3	21	16
Vocab	Class 1	29**	23	9	31**	30	24
	Class 2	15	45**	8	3	33	21
Grammar	Class 1			23	28	7	19
	Class 2			27	32	31**	30**
Listening	Class 1		27	12	16	22**	19
	Class 2		30	26**	16	11	20
Spelling	Class 1				20	12	16
	Class 2				24	12	18

material. This rapidity, of course, is a result of simply converting an already tested curriculum and methodology from teacher and classroom delivery to CALL delivery. None of the more obvious factors such as self-paced learning, the opportunity to find and correct errors over and over, or the opportunity to speak out, use new language for communicative purposes and modify interaction in the process, seem to have had much effect. As the measures focussed on the more controlled aspects of language learning, however, the results by no means suggest that language learning can be left to the computer. While the students made significant progress in most areas most of the time whether in the class or CALL group, they did not fully master the material, even in the more easily defined measures such as vocabulary.

Average scores after studying were as low as 43% in early units and rose to 59% by the end, with an average 20% improvement per unit. This indicates that the material, though written with the Japanese market in mind, was a little too difficult for these students. This may be in part because Japanese demographic trends mean that more students of lower ability enter universities each year. The important point is that the 20% improvement could have been gained by CALL before the students entered the classroom. This would almost certainly have allowed more productive recycling of the new material and consequently greater learning. The pre and post test for the whole period also showed no significant differences.

3.2 The questionnaire

While the outcomes measured revealed great similarity in the two methods of delivery, the questionnaire revealed important differences. Answers were originally on a five-point

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

Table 3. *Comparative improvement of groups – statistics (C1 = class 1; C2=class 2, the CALL group; U1-5=Units 1-5)*

		M	SD	N	t	df	P
Phrases							
U1	C1	2.71	4.08	48	–0.60	46	0.555
	C2	3.37	3.56				
U2	C1	3.08	2.48	51	–1.06	49	0.293
	C2	3.78	2.19				
U3	C1	1.21	2.00	50	0.36	48	0.723
	C2	0.96	2.79				
U4	C1	–0.1	1.36	49	–0.93	47	0.354
	C2	0.37	2.22				
U5	C1	2.35	3.01	51	–0.274	49	0.785
	C2	2.57	2.81				
Vocab							
U1	C1	2.95	2.16	48	0.01	46	0.034**
	C2	1.56	2.24				
U2	C1	4.71	4.86	51	–3.43	49	0.001**
	C2	9.04	4.15				
U3	C1	0.92	2.04	50	0.19	48	0.906
	C2	0.85	2.15				
U4	C1	3.59	2.59	49	2.22	47	0.031**
	C2	2	2.42				
U5	C1	3	2.07	51	–0.62	49	0.535
	C2	3.39	2.36				
Grammar							
U3	C1	2.33	2.35	50	–0.627	48	0.533
	C2	2.73	2.13				
U4	C1	3.73	2.75	49	–0.947	47	0.348
	C2	4.48	2.79				
U5	C1	0.74	2.36	51	–3.234	49	0.002**
	C2	3.14	2.85				
Listening							
U2	C1	2.96	2.48	51	–0.53	49	0.596
	C2	3.3	2.05				
U3	C1	1.75	3.71	50	–2.18	48	0.034**
	C2	3.85	3.08				
U4	C1	3	4.09	49	–0.16	47	0.870
	C2	3.17	2.98				
U5	C1	4.85	3.50	51	2.63	49	0.011**
	C2	2.25	3.47				
Spelling							
U4	C1	1.95	2.01	49	–0.70	47	0.489
	C2	2.44	2.75				
U5	C1	1.17	2.41	51	–0.07	49	0.947
	C2	1.21	1.95				

In the comparison, different measures of listening, dictation or comprehension, were collated into one figure. Subjects not present for both measures and classes of a unit were excluded. This made it difficult to compare totals for categories across all units as numbers would be considerably reduced.

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

Likert scale, but all positive answers and all negative answers were respectively collapsed into one category giving a three-point scale. The most salient point from the questionnaire is the more acute lack of sense of progress felt by the CALL group. While this might be justified in the case of speaking, it has no justification in terms of differences in outcomes in other categories. However, it is possible to see the greater dissatisfaction with progress in phrases and grammar as related to the lack of opportunities to utilise such new knowledge as the students had acquired. While the longer aim of this project is to integrate CALL with classroom, the reaction evidenced here means that a simple series of exercises, even when moderately varied and progressively structured within each unit, are not likely to prove sufficient to keep students motivated over the longer period unless integrated. Some support for this comes from an extension to the questionnaire given to the CALL group. Results are shown in Figure 2 and Table 4. Students enjoyed using the computer and thought it a good way to learn. They were not particularly satisfied with the materials but did not have any desire to return to using a textbook as an alternative. They did, however, have a desire to combine the two means of delivery

4 Conclusion

Pederson (1988:108) indicates that we need to do comparative work that provides *theoretical, explanatory* and *practical* information. On a theoretical level, the investigation carried out here reveals no significant advantage for students who have substantial opportunities to modify output and input in a communicative task, at least on the measures taken. It finds some support for the importance of making linguistic features salient. However, it would be necessary to control for amount of repetition or modification of input, and for opportunities to notice errors, to decide whether saliency in the form of a cloze or in the form of error messages in typed answers was the more decisive

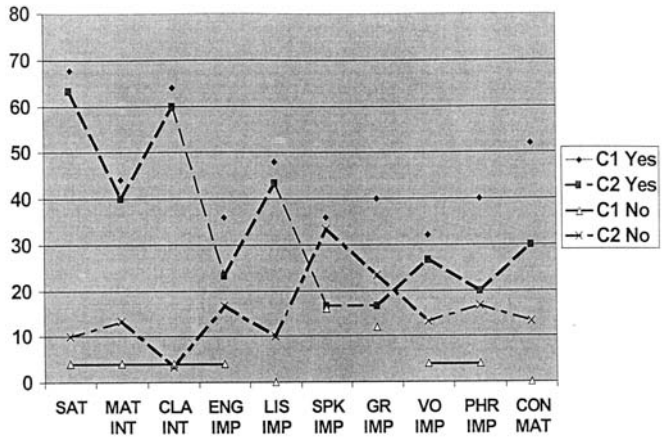


Fig. 2. Student satisfaction. (SAT = satisfied; MAT INT =found material interesting; CLA INT = found class interesting; ENG IMP = English improved; LIS IMP = listening improved; SPK IMP = speaking improved; GR IMP = grammar improved; VO IMP = vocabulary improved; PHR IMP = phrases improved; CON MAT = continue with this material. Questionnaire given in Japanese.

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

Table 4. *Feelings about using CALL*

	Yes	No
Like using computers	67	0
It's a good way to learn	73	7
Would prefer to use a textbook	17	33
Would like to use less computer	10	50
Would like to mix regular class and textbook	50	10
Questionnaire given in Japanese		

factor in the somewhat better grammatical accuracy of the CALL group. It offers support for the claim that methodology is critical and media not. From a pedagogic point of view, it suggests that a high rate of feedback combined with a high level of output modification in written form has little comparative effect. More surprisingly, it appears that the chance to have complete control over listening input has no effect on improving listening skills. Equally surprising, it suggests that the repeated typing in of words in the CALL group leads to no significant improvement in spelling. It is also clear that the criteria derived from SLA are not sufficiently clearly defined, and for that reason it is hard to decide if a particular activity in a CALL class meets a particular SLA criteria. There is a need for more attempts to define and apply these criteria. On a practical level, the results suggest that CALL can be trusted to do some kinds of work as well as a teacher, provided the methodology is sound. However, motivational factors indicate that CALL should not be relied on solely to deliver materials over longer periods. The investigation also suggests further lines of research on effects of feedback, similar to that of Nagata (1993), perhaps combining her categories with categories and results derived from classroom SLA work such as that in Lightbown and Spada (1999:103–107;141–153). There is a clear need to investigate why controlled listening is as effective as self-paced. There is a further need to understand at what stage production of language with negotiation of meaning to modify output and input works most effectively in the instructional delivery process.

The question remains whether comparative research needs to be carried out. Yildiz and Atkins (1993) give an example of what they consider a well designed CAI research project. While the results allow them to make claims about the relationship between instructional design and type of learner, they do not allow any specific claim about CAI nor about whether the presentation through CAI is more effective, or, as Chapelle (2001:26) indicated ‘value added’. In the case of the project reported here, not only do the outcomes suggest further useful research tasks, but also they put them in a comparative framework that assists educators in deciding how to integrate CALL. It is suggested that this further research could also be comparative. In addition, pursuing such research in a ‘real’ as opposed to a purely experimental setting may allow others to feel confident that the results are relatively robust. The writer believes this is valuable for two reasons. First, the teacher and classroom provide the best known yardstick by which to judge other means. Secondly, it is important to distinguish at what stage the effectiveness of the human delivery of instruction is crucial, and to show as clearly when that is the case

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

as when it is not. This research has obvious weaknesses. For example, the software was nor able to store data on student activity. Such data would have allowed a much more comprehensive analysis. However, in terms of clarity of description, use of both judgmental criteria related to SLA, and empirical outcome measurements, and in control of variables, the writer hopes that some response has been made to earlier criticisms of such research, and some case made for its continuation.

References

Chapelle, C. A. (2001) *Computer Applications in Second Language Acquisition*. Cambridge: CUP.

Chapelle, C. A. (1998a) Multimedia CALL: lessons to be learned from research on instructed SLA. *Language Learning and Technology* **2**(1):22–34.

Chapelle, C. A. (1997) Call in the year 2000: still in search of search paradigms? *Language Learning and Technology* **1**(1):19–43.

Chapelle, C. A. and Jamieson, J. (1986) Computer-assisted language learning as a predictor of success in acquiring English as a second language. *TESOL Quarterly* **20**(1):27–46.

Clark, R. E. (1994b) Media and method. *Educational Technology Research and Development* **42**(3):7–10.

Clark, R. E. (1994a). Media will never influence learning. *Educational Technology Research and Development* **42**(2):21–29.

Clark, R. E. (1985c) Evidence for confounding in computer-based instruction studies: Analysing the meta analyses. *Educational Communication and Technology Journal* **33**(4).

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

Dunkel, P. (1991) Research on the effectiveness of computer-assisted instruction and computer-assisted language learning. In: Dunkel P. (ed.) *Computer-Assisted Language Learning and Testing*. New York: Newbury House.

Levy, M. (2001) Coherence and direction in CALL research: comparative designs. In: Cameron K. (ed.), *C.A.L.L. – The Challenge of Change*. Exeter: Elm Publications, 5–13.

Levy, M. (1997) *Computer Assisted Language Learning*. Oxford: Clarendon Press.

Lightbown, P. and Spada, N. (1999) *How Languages are Learned*. Oxford: OUP

Nagata, N. (1993) Intelligent computer feedback for second language instruction. *The Modern Language Journal* **77**(3):330–339

Pederson, K. M. (1988). Research on CALL. In: Smith W. F. (ed.), *op. cit.*, 99–132.

Pederson, K. M. (1986) An experiment in computer-assisted second language reading. *The Modern Language Journal* **70**(1):36–41

Schrupp, D. M., Busch, M. D. and Mueller, G. A. (1983). Klavier im haus – an interactive experiment in foreign language instruction. *CALICO Journal* **1**(2):17–21.

Yildiz, R. and Atkins, M. Evaluating multimedia applications. *Computers in Education* **21**(1/2): 133–139.

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

Appendix 1

Comparison of instruction of one unit through the teacher and CALL
(T = task; I = implementation; F = feedback)

Directions, understanding and giving, and store names

	CLASSROOM	**CALL**
Aim	*Introduce new phrases related to direction giving*	
T1a	Match word or phrase with picture representing it. e.g. turn right, it's across from	
I	Put letter beside phrase in box next to matching picture	Drag and drop phrase to picture. Phrases in boxes on right of screen, pictures on left.
F	Student call out. Teacher correction & comment.	Students check answer as like by pressing 'Check' button. Wrong answers return to original position. Feedback = right/wrong
1b	Translate above phrases into Japanese.	
T1b	Call out in Japanese	Drag and drop Japanese phrase to English phrase.
F	Student call out. Teacher correction & comment.	Checking and feedback as for T1
Aim	*Introduce vocab for stores*	
T2 & I/F	As for T1a & b but with names of stores. e.g. stationery store	As for T1a & b but with names of stores. e.g. stationery store
Aim	*Listening comprehension of new vocab & phrases in context.Introduce functional phrases indirectly*	
T4a	Listen to 4 sets of directions dialogues to 4 stores while looking at map and locate goal.	
I	Mark store on map by writing name. on building e.g. stationery store	Each building has a letter (e.g. a, b, c) imposed on it. Type in the letter of the building where the store is located and the name of the store.
I	CD played twice straight through all 4 directions: repeated for each direction twice with possible student response & teacher feedback after each one. After second time, give answer and replay once more, pausing at key phrases to speak out and repeat. Do for each section. Finally, replay all 4 straight through.	Listen freely Clues to help
F	Call out answer if can after twice straight through. Feedback possible after each playing of each section; confirmation on third play through with teacher repetition of key phrases.	Check anytime. Feedback = right/wrong
T4b	Listen to new directions (not dialogue) looking at same map. Locate and name store.	

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

	CLASSROOM	CALL
I/F	As for T3	As for T3
Aim	*Relate new phrases & vocab to own experience while beginning production, and having opportunity to check output*	
T5	Listen to and answer three questions using new structures/vocab to convey personal information e.g. 'What's across from your home?'	
T5a	Catch question accurately	
I	Dictate each question	
I	All 3 played through once,then each questionthree times with a pause between each, then all once again.	Listen freely
I	Write down question	Type in question
F	Student call out. Teacher correction & comment. Write on board.	Check freely. Feedback shows which part of the answer is correct i.e. which words or part words in own answer are part of correct answer.
T5b	Speak out with partner – ask and answer.	Write out question and answer on worksheet and hand in.
F	Teacher check as walk round. Selected pairs speak out. Correction, comments. Write on board if necessary.	Corrected and returned with general comments following week.
Aim	*Introduce more functional phrases (e.g. how do I get to), practise controlled production of these plus previously learnt phrases and vocab*	
T6	Listen to gap fill of directions dialogue.	
I	Fill in gaps using list of items provided.	Type in blanks. Clues provided.
I	Play whole dialogue once. Play sentence by sentence, with pauses, 3 times. Play whole dialogue once.	Listen freely
F	Pair read out. Teacher correction, comment.	Right/wrong
I	Speak out dialogue with partner putting in provided substitutions. Selected pairs speak out. Make dialogue based on own experience. Speak out.	Write out dialogue using substitutions. Write out using own experience. Hand in.
F	Teacher check as walk round. Pair read out. Correction, comment.	Corrected and returned with general comments following week.
Aim	*Exchange information with a partner using new functional phrases and vocabulary*	

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

	CLASSROOM	**CALL**
T7	Information gap. Looking at same map but with different information marked in, ask and answer how to get to selected places.	Information gap. Looking at map, ask and answer how to get to selected places.
I	Ask and answer, writing down answer on respective versions of map. Table of typical directions provided above map to help.	Cloze exercise. Looking at map read questions and directions to specified places and fill in blanks. Table of typical directions provided above map to help. Clues to help.
F	Selected pairs speak out. Correction as needed.	Feedback = write/wrong
Aim	*Check knowledge of directions*	
T8	Looking at new map, read and follow directions.	
I	Write down number of building you get to.	Type in number of building you get to.
F	Call out answer. Correction, comment.	Right/wrong
Aim	*Check knowledge of store vocab*	
T9	Name store that matches the definition.	
I	Write store name next to definition	Type in store name next to definition
F	Call out answers. Correction, comment	Check anytime. Feedback tells which part of typed in word is correct.
Aim	*Check listening ability for directions*	
T10	Listen to invitation with directions	
	Gap fill – choose from list of phrases.	Gap fill.
I	Write in blanks.	Type in. Clues available.
I	Play CD four times.	Play CD freely
F	At end student read out. Correction, comment. Play CD again.	Check freely. Feedback = right/wrong

Appendix 2

Translations / answers in brackets added.

Mini Test 5

I. ..... (Translate into Japanese)

1. theater
2. hospital
3. video shop

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press

4. appliance store
5. stationery store

II. ..... (Translate the underlined parts into English)

1. ..... (across from the post office)
2. ..... (on this street)
3. ..... (on the corner of Ginza Street and Namiki Street)
4. ..... (past the bus stop)
5. ..... (between the restaurant and the camera shop)

III. ..... (Translate into Japanese)

1. How do I get to the bank?
2. Go straight down this street.
3. Turn left.
4. Is there a post office near here?

IV. ..... (Write down the words the teacher speaks out)

V. ..... (Dictate the conversation you hear on the CD played by the teacher)

VI. ..... (Listen to the conversation and answer the questions below)

Conversation 1

1. What does the woman want to buy?
2. What kind of shop is she looking for?
3. Where do you turn right?
4. What's on the corner of Pine and Third?
5. What' across from the shop she is looking for?

https://doi.org/10.1017/S0958344002001210 Published online by Cambridge University Press