*ReCALL* **25**(2): 165–198. 2013 © European Association for Computer Assisted Language Learning
doi:10.1017/S0958344013000013
First published online 27 March 2013


# *A meta-analysis of effectiveness studies on computer technology-supported language learning*

MAJA GRGUROVIĆ
*Department of Linguistics, University of Illinois at Chicago, 1710 University Hall, 601 S. Morgan St., Chicago, IL, 60607, USA*
*(email: maja@uic.edu)*

CAROL A. CHAPELLE
*Department of English, Iowa State University, 339 Ross Hall, Ames, IA 50011, USA*
*(email: carolc@iastate.edu)*

MACK C. SHELLEY
*Departments of Statistics and Political Science, Iowa State University, 1413 Snedecor, 539 Ross Hall, Ames, IA 50011, USA*
*(email: mshelley@iastate.edu)*

**Abstract**

With the aim of summarizing years of research comparing pedagogies for second/foreign language teaching supported with computer technology and pedagogy not-supported by computer technology, a meta-analysis was conducted of empirical research investigating language outcomes. Thirty-seven studies yielding 52 effect sizes were included, following a search of literature from 1970 to 2006 and screening of studies based on stated criteria. The differences in research designs required subdivision of studies, but overall results favored the technology-supported pedagogy, with a small, but positive and statistically significant effect size. Second/foreign language instruction supported by computer technology was found to be at least as effective as instruction without technology, and in studies using rigorous research designs the CALL groups outperformed the non-CALL groups. The analyses of instructional conditions, characteristics of participants, and conditions of the research design did not provide reliable results because of the small number of effect sizes representing each group. The meta-analysis results provide an empirically-based response to the questions of whether or not technology-supported pedagogies enhance language learning, and the process of conducting the meta-analysis pointed to areas in research methodology that would benefit from attention in future research.

Keywords: research methods, meta-analysis, secondary research, research synthesis, second/foreign language learning, computer-assisted language learning.

## 1 Introduction

Teachers and researchers in computer-assisted language learning (CALL) are frequently asked by their colleagues and policy makers about the value of technology for second language learning relative to the classroom instruction that they see as traditional. Professionals in CALL often find this comparison question frustrating. Recognizing the complexity of the question, CALL researchers have argued that technology and language learning cannot adequately be studied through research comparing technology-supported learning with traditional classroom pedagogy (Chapelle, 2003; Dunkel, 1991; Pederson, 1987). Instead, CALL specialists aim to create ideal language learning conditions through strategic use of pedagogies developed around interactive video, learner-computer interactions, corrective feedback, tasks with linguistic support, and intercultural communication, for example. They seek evidence for the effects of these innovations on learners' interactions, attitudes, and outcomes; and they design research to be informative to the community of specialists in CALL (Chapelle, 2007).

Policy makers and many language teachers remain interested in the comparative question: What has the research shown about the comparison between classes in which CALL is used and those in which computer technology is not used for language learning? It is not the question that the large majority of CALL specialists ask, but in a political sense, it would be useful if CALL specialists could answer it. Indeed, over the past decades, studies have been carried out to address precisely the policy-oriented question that remains of interest to some. If one were to examine all of those studies taking the researchers' original conceptualization of the CALL and traditional condition, would an overall favorable effect be found for the CALL condition?

The research methodology of meta-analysis offers the analytical tools for addressing such a question in a principled way. Meta-analysis, the best described and most common form of research synthesis (Norris & Ortega, 2006; Lipsey & Wilson, 2001), includes quantitative primary studies which report descriptive or inferential statistics. Meta-analysis uses the effect size to summarize results so that each finding is expressed as a standard unit called the effect size.[1] An effect size gives a direct measure of the impact of an intervention in terms of how much difference is found between groups or points in time relative to the standard deviation of the difference. Effect sizes are not affected by the sample size, unlike the results of tests of statistical significance, in which for large samples even a small difference may be statistically significant. Thus, the result of a study expressed as an effect size can be more meaningful and more interpretable than whether or not the result is statistically significant, particularly when it is important to know the magnitude of an intervention.

Meta-analysis is suited not only to investigating specific variables such as corrective feedback (Li, 2010) but also broader policy questions in the educational policy arena (Oswald & Plonsky, 2010). For example, research sponsored by the US Department of Education (2009) conducted exactly such a meta-analysis comparing

[1] Definitions of statistical terms are provided in the glossary at the end of the paper.

technology vs. traditional classroom conditions on student learning across subject areas from 1996 through 2006. Their result indicating more positive outcomes from the technology groups is precisely the level of information of interest to policy makers.

## 2 Meta-analyses of research on technology for language learning

Interest in the comparative question about technology use and meta-analysis of these studies has been evident for many years in education (e.g., Kulik & Kulik, 1987, 1991). One such meta-analysis of the effectiveness of instructional technology in higher education included studies on the effectiveness of computer use in science, mathematics and language teaching (Kulik, 2003). Forty-six studies that Kulik (2003) located, including seven CALL studies published between 1992 and 1998, were of appropriate methodological quality. The results showed that CALL had an overall positive instructional effect with five studies having educationally meaningful effect sizes larger than + 0.25. In addition, the median effect of CALL programs in all seven studies was an increase of language test scores of 0.6 of a standard deviation which indicates a moderate to large improvement in student performance.

Despite the appearance of some CALL studies in this meta-analysis, as Felix (2005a) pointed out, "the surprisingly scarce meta-research specifically related to CALL tells us very little about actual or potential effectiveness of the use of ICT in second language learning" (*op. cit.*:284) from a quantitative, outcomes-oriented perspective. Felix's search for such meta-analyses found several on first-language instruction, but only one that had addressed specifically the question of CALL effectiveness. Zhao's (2003) meta-analysis included nine empirical CALL studies published in five journals in the five-year period 1997–2001. Each study measured language learning outcomes that could be attributed to some kind of technology (although not necessarily computer technology). An effect size of + 1.12 was found, which indicates a large positive effect of technology use.

The positive effects on language learning associated with technology, found in Kulik (2003) and Zhao (2003), also appear in some research syntheses producing narrative findings. Zhao's (2003) synthesis of 156 journal articles by aspects of language learning concluded that technology can be effective in almost all areas of language education. Liu, Moore, Graham, and Lee (2002) looked at computer use in second/foreign language learning from 1990 to 2000 and Felix's (2005b) analysis of CALL effectiveness research included the period from 2000 to 2004. In addition to these meta-analyses targeting the overall question of CALL effectiveness, a few meta-analyses investigated the effects of specific features of CALL such as effects of computer-mediated glosses on reading comprehension and vocabulary learning (Abraham, 2008) and computer-mediated L1 glosses on reading comprehension (Taylor, 2006). In the two meta-analyses targeting specific features of CALL, the results indicated clear, positive results for the computer-mediated conditions.

These meta-analyses and research reviews offer snapshots of aspects of the overall picture emerging over the past years, but they do not perfectly satisfy the need for a straightforward answer to the policy question. To answer the question, a systematic meta-analysis of the entire body of research using well-documented inclusion criteria is needed. Solid guidance and examples for meta-analytic research as one type of

synthetic methodology in applied linguistics are provided by Norris and Ortega (2006). Norris and Ortega (2006: 4) define research synthesis as "the systematic secondary review of accumulated primary research studies." They present the following defining features of a research synthesis (*op. cit.*: 6–7):

1) It clearly states how primary literature was searched and how the selection of studies was done. This means that before the authors go into reviewing primary studies, they need to explain how they obtained them and which criteria were used.
2) It concentrates on the variables and data in primary research more than conclusions drawn by their authors. Instead of just using the findings of primary researchers, secondary researchers look at the data themselves (to the extent that is possible) and make conclusions.
3) It presents conclusions by looking at categories of data and methodologies that cut across studies. To be able to synthesize findings from a number of studies that have a number of variables, one needs to establish super-ordinate categories that are going to encompass all of the studies.

The meta-analysis reported here was guided by this three-point definition of research synthesis. It sought to respond to the following research questions which aimed to yield an overall result as well as to isolate any factors that may play a role in effectiveness:

1) Is pedagogy supported by computer technology effective in promoting second/foreign language development relative to pedagogy that does not include computer technology?
2) Are results affected by instructional conditions used in the study – type of technology, degree of integration, length of instructional treatment?
3) Are results affected by the characteristics of the participants used in the study – language proficiency level, or native language?
4) Are results affected by the conditions of the research design – the setting where the study took place, the language taught, the number of participants, and the method of assignment into groups?

## 3 Methodology

### *3.1 Study identification and retrieval*

The first step in the preparation for the research synthesis and meta-analysis was to conduct an extensive literature search for studies comparing language instruction with computer technology and instruction without computer technology (here the latter is called traditional instruction). The literature search covered the period from 1970 to the end of 2006 and included a computer search of three electronic databases and manual search of six journals.

The databases searched were: Linguistics and Language Behavior Abstracts (LLBA), Education Resources Information Center (ERIC), and Dissertation Abstracts (DA). LLBA and ERIC were searched through Cambridge Scientific Abstracts (CSA Illumina) and DA were searched through ProQuest Digital Dissertations. The following terms in

various combinations were used in the search: computer assisted language learning, computer assisted instruction, computer assisted learning, computer based instruction, computer instruction, traditional instruction, second language, foreign language, differ*, compar*, PLATO, and TICCIT. The abstracts which resulted from searching key terms were carefully read through and full articles skimmed whenever they were available.

A manual search of the following major CALL journals was conducted: *Computer Assisted Language Learning* (*CALL*), *System*, *CALICO Journal*, *ReCALL*, *Language Learning and Technology* (*LL&T*), as well as *TESOL Quarterly* to make sure no articles were missed. This search of the journals revealed some twenty articles other than those found through the database search because some of the journals were not abstracted in the databases (for example *LL&T*) and some comparative studies did not contain the key words used in the search. Overall, more than 200 sources were identified in computer and manual searches.

The database search results included more than twenty-five unpublished sources, mainly doctoral dissertations and reports, which were kept in the original pool of studies. Although claims might be made that published studies are of better methodological quality than unpublished literature because the former has been through a peer review process (as cited in Norris & Ortega, 2006: 20), we decided to include unpublished literature for several reasons. First, we believe that unpublished literature helps paint a more exhaustive picture of CALL effectiveness because it increases the number of sources looked at and helps avoid publication bias, i.e., the tendency to publish only those studies with statistically significant findings. Second, unpublished work is on average lengthier and more in-depth than published journal articles (because it does not have to conform to length restrictions of most journals) which in turn provides details necessary for a comprehensive research synthesis. Third, authorities in meta-analytic research methodologies, Norris and Ortega (2006: 22), argue that unpublished literature should be included whenever possible.

Although we made sure that all of the unpublished literature that came up in the electronic search was included, we are aware that this search could not have covered all of the unpublished CALL comparison work and we acknowledge this limitation. The types of unpublished sources we included when they came up in the search were electronic copies of doctoral dissertations and reports that could be obtained from the three databases listed above or through the interlibrary loan service. Doctoral dissertations which were available only for purchase from UMI, Ann Arbor, MI were not included due to the lack of financial resources available for this project.

### *3.2 Study eligibility criteria and coding*

Out of more than 200 studies retrieved, 85 studies met the research synthesis criteria and were included in the synthesis. Appendix A (appears online) contains inclusion and exclusion criteria that were used.

Once studies were retrieved, one of the authors coded them. Eighteen coding criteria were used: Primary skill, Secondary skill, Language(s) taught, Participants' language level, Number of participants, Participants' native language, Setting, Technology used, Technology details, Variable(s), Description of groups/courses,

Type of CALL aspect, Length of CALL aspect, Research design, Results, Results details, Pedagogy, and SLA aspect. The definition of each criterion and the descriptors are given in Appendix B1 (appears online) together with the list of original studies (Appendix B2 online).

After all the studies were coded, they were divided by primary skills and a research synthesis was written for each of seven skills/ knowledge areas (reading, writing, vocabulary, communication, grammar, pronunciation, integrated skills). Because of the difference in the number of studies that appeared in each of these categories, we did not attempt a quantitative analysis that would separate out effects based on these skill areas. Instead, the research synthesis describing patterns in these data was written. In addition, a database of all comparison studies can be accessed online and can be searched using the above criteria or user-supplied keywords.

The next step involved separating studies for the meta-analysis. Since in the meta-analysis effect sizes can be calculated only if the original study reported descriptive or inferential statistics as needed in each context, studies that did not report statistics or those that reported insufficient test results, as well as qualitative research studies, were excluded. The study was included if it:

1) Measured participants' performance on language tests.
2) Used an experimental[2] or quasi-experimental design.
3) Employed a pre-test/post-test design or post-test design only.

The study was excluded if it:

1) Measured factors other than language learning outcomes (for example attitudes, motivation, study skills, participation).
2) Reported results of tests given during the treatment rather than at the end.
3) Did not report statistics or reported statistics that were insufficient to calculate the effect size.

The last criterion was the main reason for exclusion of studies. The statistics necessary to calculate the effect size were the mean[3], standard deviation, and sample sizes but a surprisingly large number of studies did not report some of these statistics, most often standard deviations. In some cases, we were able to utilize other statistics (t values, F values, and ANCOVAs) or raw data to calculate desired parameters. If no other statistic was available but the effect size was reported, we used the authors' reported effect sizes values. The list of excluded studies together with the reasons why they were excluded from the meta-analysis can be found in Appendix C (online).

Once all the studies for the inclusion in the meta-analysis were identified, an additional coding of studies was performed by the primary coder, this time focusing in detail on methodologies used in the study. A coding form was developed based on the sample coding sheet in Cooper (1998: 32–35). Additionally, methods sections of meta-analytic studies in Norris and Ortega (2006) were consulted for best coding practices and examples of coding categories and forms. Secondary coding was performed by

[2] Definitions of research designs are provided in the glossary at the end of the paper.
[3] Definitions of statistical terms are provided in the glossary at the end of the paper.

the primary coder and two other independent coders, doctoral students in applied linguistics. A coder training session was conducted on two sources (Tozcy & Coady, 2004; Nagata, 1996) and results were compared among coders. Then, the coders were provided with a coding form and manual (see Appendix D online).

In the second coder training session, each independent coder analyzed three studies and compared the results with those of the primary coder. Discrepancies were discussed and questions clarified. Next, the rest of the studies were divided between the two secondary coders. After all the coding was performed, the intercoder reliability was calculated between the primary coder and each of the secondary coders using simple agreement calculations (see Norris & Ortega, 2006: 26). The intercoder reliability between the primary coder and the first secondary coder was 89% and between the primary coder and the second secondary coder 87%. Overall, intercoder reliability was 88% over 23 categories (see Appendix D online, Part 2–3, items 10–32).

In the second round of coding, 37 studies were coded and included in the meta-analysis (see Appendix E for the list). The studies appeared between 1984 and 2006. All effect sizes per study were included. In studies with multiple dependent variables (for example, measures of reading, listening, and grammar) and with multiple comparison groups (for example, more than one experimental group), it was possible for one study to produce more than one effect size. The averaging of effect sizes was not done in order to keep the actual variation in outcomes that had been found in the studies. Thirty-seven studies in the meta-analysis produced a total of 144 effect sizes.

### *3.3 Analysis*

The first step in the analysis involved examining whether experimental and control groups were equivalent at pre-test. For each of the effect sizes, it was necessary to determine this prior equivalence of groups which indicates that the groups are equal at the outset of the study and that, therefore, the difference at post-tests can be attributed to the treatment. Primary authors measured the equivalence of groups with tests of statistical significance. Our examination revealed four different Groups based on whether pre-tests were administered and whether or not the pre-test scores were significantly different (see Table 1).

*3.3.1 Group 1: Equivalence of groups at pre-test found.* In Group 1, equivalence of groups was found at pre-test indicating that there was no statistically significant difference between experimental and control group at pre-test for the given variable. In this case, the effect size of interest for the meta-analysis was calculated using the standardized mean difference statistic which takes into account the means of control and experimental groups at post-test, their standard deviations, and sample sizes. The formula used is:

$$ES = \frac{\bar{X}_{G1} - \bar{X}_{G2}}{s_p} \text{ (source Lipsey \& Wilson, 2001: 48)}$$

$$Sp = \sqrt{\frac{(n_1 - 1)s_1{}^2 + (n_2 - 1)s_2{}^2}{n_1 + n_2 - 2}} \text{ (source Lipsey \& Wilson, 2001: 198)}$$

Table 1 *Four Groups used in the analysis*

| Group | Criteria | Contrast | Effect size statistics | Number of studies (K) | Number of ES |
|---|---|---|---|---|---|
| 1. | Equivalence of groups found at pretest | Experimental-control at posttest | Standardized mean difference | 14 | 32 |
| 2. | Equivalence of groups not tested (without pretests) | Experimental-control at posttest | Standardized mean difference | 14 | 81 |
| 3. | Equivalence of groups not found at pretest | Pre and posttest for experimental group | Standardized mean gain | 9 | 15 |
| 4. | Equivalence of groups not tested (with pretests) | Pre and posttest for experimental group | Standardized mean gain | 5 | 16 |

$\overline{X}_{G1}$ is the mean of group 1, $\overline{X}_{G2}$ is the mean of group 2, $s_p$ is the pooled standard deviation, $S_1$ is the standard deviation of group 1, $S_2$ is the standard deviation of group 2, $n_1$ is the number of subjects in group 1, and $n_2$ is the number of subjects in group 2. The contrast is based on the experimental and control groups at post-test. This Group has 14 studies and 32 effect sizes.

*3.3.2 Group 2: Equivalence of groups at pre-test unknown (no pre-tests).* In Group 2, the equivalence of groups was not tested because the primary authors did not administer pre-tests for the given dependent variable. This was very common in cases of random assignment of participants since the authors assumed the groups were the same given that their participants were randomly chosen. In cases when assignment of participants to groups was non-random as with intact classes, some primary authors used measures of other variables to establish equivalence of group. For example, Chenoweth, Ushida, and Murday (2006) used students' SAT scores as well as background and technology questionnaires. In this group, the contrast was experimental-control group at post-test and the effect size statistic standardized mean difference (see the formulas above). In cases when primary researchers did not report some of the variables necessary to calculate effect sizes, we obtained the effect size statistics from F and t values using formulas in Lipsey and Wilson (2001), from Hedges' g value, or we used the ES values (Cohen's d) already calculated by primary researchers. This Group contains 14 studies, which produced 81 effect sizes.

*3.3.3 Group 3: Equivalence of groups at pre-test not found.* For studies in this Group, the equivalence of groups was tested on the pre-tests but was not found. Some of the authors introduced covariates to level out this difference between groups (for example Payne & Whitney, 2002; Al-Jarf, 2002; Odenthal, 1992). Since the groups were not equal at pre-test, we decided to look at the improvement made by

the experimental group from pre- to post-test and used the standardized mean gain effect size formula:

$$ES = \frac{\bar{X}_{T2} - \bar{X}_{T1}}{s_p} \text{ (source Lipsey \& Wilson, 2001 : 44)}$$

$$Sp = \frac{S_1 + S_2}{2} \text{ (source Rawdon, Sharp, Shelley, \& Thomas, 2012; Shelley, Rawdon, Sharp, Thomas, \& Schalinske, 2007)}$$

$\overline{X}_{T2}$ is the mean at post-test, $\overline{X}_{T1}$ is the mean at pre-test, $s_p$ is the pooled standard deviation, $S_1$ is the standard deviation at pre-test, and $S_2$ is the standard deviation at post-test. In cases when some of these values were not reported, we used t values or already calculated Cohen's d values. This Group contains 9 studies and 15 effect sizes.

*3.3.4 Group 4: Equivalence of groups at pre-test unknown (with pre-tests).* In the fourth Group, equivalence was not tested although there were pre-tests. As with Group 3, pre- and post-test contrast for the experimental group was used and standardized mean gain effect size value calculated (see the formula above). This Group has 5 studies and 16 effect sizes.

Since effect sizes tend to be upwardly biased when based on small samples, particularly those less than 20 (Lipsey & Wilson, 2001: 48), the adjustment of effect sizes was performed for all four groups. Hedges' correction formula (Lipsey & Wilson, 2001), which takes into account the sample size and multiplies it by the original ES value, was used:

$$ES' = \left[1 - \frac{3}{4N - 9}\right] ES \text{ (source Lipsey \& Wilson, 2001 : 49)}$$

$ES'$ is the corrected (unbiased) effect size, $N$ is the total sample size $n_1 + n_2$ and $ES$ is the biased standardized mean difference/gain (see formulas above). Furthermore, the $ES'$ values were used to calculate the mean $E\overline{S}$ value for each of four groups. The mean $E\overline{S}$ per group were found using the following formula:

$$E\bar{S} = \frac{\sum WES'}{\sum W} \text{ (source Lipsey \& Wilson, 2001 : 132)}$$

Where $W$ is the inverse variance weight and $E\overline{S}$ is the weighted effect size. In this case, $W$ was calculated from the standard error using the formulas in Lipsey and Wilson (2001: 49). In order to expedite the calculation process, the macro created by Lipsey and Wilson (2001: 208–220) was used with the SPSS data file. The weighted mean $E\overline{S}$ for each of the groups, together with confidence intervals and p and Q values, are presented in Table 4. Q values (homogeneity) indicate the dispersion of effect sizes around the mean and whether they create a normal distribution; a normal distribution means that there is no sampling error and that effect sizes are independent from each other.

### 3.4 Interpreting effect sizes

To interpret the value of the standardized mean difference effect size (Groups 1 and 2), Cohen's (1988) guidelines were used. According to Cohen, small effect sizes are less or equal 0.2, medium are around 0.5, and large effect sizes are equal to or greater than 0.8. However, these guidelines cannot be used for interpretation of the magnitude of standardized mean gain effect size (Norris & Ortega, 2006) in Groups 3 and 4 because pre-post test contrasts tend to be larger than experimental-control contrasts so the results were compared to previous research that used the same mean gain statistic. In particular, we compared our findings to Norris and Ortega (2000) and Jeon and Kaya (2006), whose effect sizes of 1.66 and 1.57 respectively were interpreted as large.

Table 2 shows that the mean effect sizes in each group are small (Groups 1, 3, and 4), or almost zero (Group 2). The *p*-values obtained demonstrate statistically significant results for Groups 1, 3, and 4, indicating that the results are not due to chance. The weighted mean effect size of 0.257 for the first Group shows that CALL groups performed better than non-CALL groups. Although the effect size is not very large, it is statistically significant. Also, there is a 95% chance that the mean ES value falls between 0.1728 and 0.3416.

With the effect size of 0.0207, the lowest mean effect size of all Groups, the differences between CALL and non-CALL groups may be non-existent for studies in Group 2. A large standard deviation is a sign of a very large variation in this group, suggesting that this research design without controlling for knowledge at the beginning of the study produces the most variable results.

Table 2 *Mean effect sizes, standard deviations, confidence intervals, p and Q values for 4 Groups*

| Group | Criteria | Weighted Mean $E\bar{S}$ | Standard Deviation (weighted) | 95% Confidence Interval | | *p* | Q |
|---|---|---|---|---|---|---|---|
| | | | | Lower | Upper | | |
| 1. | Equivalence of groups found at pretest **K = 32** | 0.2572 | 0.543 | 0.1728 | 0.3416 | <0.001 | 158.7555* |
| 2. | Equivalence of groups not tested (without pretests) **K = 81** | 0.0207 | 0.79 | −0.0657 | 0.1071 | 0.639 | 321.4127 |
| 3. | Equivalence of groups not found at pretest **K = 15** | 0.3291 | 0.318 | 0.2522 | 0.4060 | <0.001 | 65.85* |
| 4. | Equivalence of groups not tested (with pretests) **K = 16** | 0.4232 | 0.735 | 0.2617 | 0.5846 | <0.001 | 79.7321* |

Note: $*p < 0.05$

The effect size of 0.3291 for Group 3 is an indicator of the average improvements made by CALL groups from pre-test to post-test. There is a 95% chance that, on average, CALL groups will improve anywhere from 0.2522 to 0.4060 during the course of the study. This result is also statistically significant at the 0.05 level.

Finally, Group 4 has the largest effect size (0.4232), indicating the largest gains from pre-test to post-test for CALL groups. Although the ES sizes here have a large standard deviation, the confidence interval is entirely positive.

The Q values for all four groups are statistically significant, and therefore the assumption of the homogeneity of the sample cannot be accepted (i.e., the differences in homogeneity are significant). This result is not surprising given that more than one effect size per study was included and that studies with more effect sizes contributed more weight than those with only one effect size.

With homogeneity not found in the sample, we returned to the original corrected effect sizes (*ES′*) and looked at possible ways to address this issue. One of the options suggested by Norris and Ortega (2006: 29) was to combine effect sizes so that one effect size per study is included. In cases when there was more than one effect size per study, an average value was used to represent that study if all effect sizes came from the same group. In cases when effect sizes came from different groups (participants in experimental and control groups from effect size 1 and those in experimental and control groups from effect size 2 were not the same people), effect sizes were not averaged.

Since statistics used to calculate effect sizes were different (standardized mean gain for Groups 1 and 2 and standardized mean difference for Groups 3 and 4), we first checked whether the combination of effect sizes from four different Groups would be possible. Therefore, we created a new variable (Group) and coded each effect size based on the Group it belonged to (1, 2, 3, or 4). Then, a general linear model was estimated to see whether the Groups were different. Values for the Scheffé and Bonferroni multiple comparisons showed statistically significant differences between Groups 2 and 4 while the Tamhane procedure (assuming unequal group variances) found, in addition to Groups 2 and 4, a significant difference between Groups 2 and 3.

This result led to the conclusion that all four Groups should not be combined, which supported our original idea that effect sizes obtained using different statistics should be analyzed separately. No significant differences were found between Groups 1 and 2 nor between Groups 3 and 4; therefore, we were able to combine effect sizes in Groups 1 and 2 as well as in Groups 3 and 4. In addition, since one study could produce several effect sizes depending on whether the contrast was between pre-test and immediate or delayed post-tests, it was decided to average them. This decision was based on the results of a *t*-test that showed there was no statistically significant difference between all immediate and delayed post-tests in the sample of 144 original effect sizes.

A mean *ES* for each unique sample was calculated for studies having more than one effect size and they were put into two clusters: 1 and 2 (standardized mean difference Group) and 3 and 4 (standardized mean gain Group). When one study had effect sizes belonging to both Groups, that study contributed two effect sizes in total. The final number of effect sizes from 37 studies was 65, with 49 studies giving effect sizes for the standardized mean difference Group and 16 for the standardized mean gain Group.

## 4 Results

The results for two groups with the weighted mean effect sizes are presented in Table 3. According to Cohen's (1988) guidelines, the mean effect size value of 0.2353 is small. This value is statistically significant at the 0.05 level. The confidence interval shows that, on average, with 95% confidence, effect sizes are estimated to range from very small (0.1435) to small to medium (0.3271). This interval is always positive, with CALL groups performing better than non-CALL groups on average. The standard deviation of 0.633 is rather large, indicating large variability in the sample, but its value has been adjusted (weighted) to take into account the inverse variance. Overall, this is a more conservative measure than an unweighted standard deviation.

As for studies in Groups 3 & 4, the effect size of 0.352 would be considered small when compared to findings of other studies (Norris & Ortega, 2006; Jeon & Kaya 2006). This result is statistically significant at the 0.05 level with the average improvements of CALL groups from 0.2601 to 0.4439 due to the CALL treatment.

### *4.1 Effectiveness for language development*

The first research question asked about the effectiveness of pedagogy supported by computer technology in promoting language development. The effectiveness of computer technology was assessed by looking at mean effect sizes values for two groups of studies (Groups 1 & 2 and 3 & 4). As shown in Table 3, language instruction with computer technology was more effective than instruction without it since CALL groups showed better performance than non-CALL groups. The mean effect size of 0.2353 is small, indicating that scores of CALL groups were 0.23 standard deviations higher than scores of non-CALL groups (interpretation from Norris & Ortega, 2006: 33). Although small, this result is not due to chance judging by the statistically significant p value. Moreover, the upper level of the confidence interval shows that scores of CALL groups can be up to 0.33 standard deviations higher than those of non-CALL groups.

The average effect size of 0.35 for Groups 3 & 4 shows improvement of CALL groups from pre- to post-test due to treatment which involved computer technology.

Table 3 *Mean effect sizes, standard deviations, confidence intervals, and p values for two Groups*

| Groups | Group name | Weighted Mean $E\bar{S}$ | Standard Deviation (weighted) | 95% Confidence Interval | | $p$ |
|---|---|---|---|---|---|---|
| | | | | Lower | Upper | |
| 1 & 2 | Standardized mean difference Group **K = 49** | 0.2353 | 0.633 | 0.1435 | 0.3271 | <0.001 |
| 3 & 4 | Standardized mean gain Group **K = 16** | 0.3520 | 0.400 | 0.2601 | 0.4439 | <0.001 |

When compared with previous literature, this effect size value is small. Although gains are not large, they are statistically significant and always positive judging by the all-positive confidence interval.

### 4.2 Instructional conditions

Since technology for language teaching can be used in many different ways, we categorized the studies in the database according to type of technology, degree of integration with the course, and length of instructional treatment.

*4.2.1 Type of technology.* For the type of technology analysis, the studies were coded for the following:

1) CALL program (a computer program originally made for language learning).
2) Computer application (a computer program not originally made for language learning, e.g., Microsoft Word).
3) CMC (computer mediated communication program that allows synchronous or asynchronous communication).
4) Web (use of authentic materials and resources on the WWW).
5) Course management system (a system for managing course content, e.g., WebCT).
6) Online course (a course delivered completely online).

In some studies more than one technology was used so additional categories were developed:

1) CMC + web + course management system.
2) CMC + course management system.
3) Computer application + CMC + web + course management system.

The results for Groups 1 & 2 presented in Table 4 show that the majority of studies used CALL programs with 31 effect sizes in this subgroup. The average effect size of 0.46 is around medium with a large standard deviation and all positive confidence interval from 0.14 (small) to 0.78 (around large). These results indicate that groups using CALL programs performed better than non-CALL groups. The effectiveness of CALL programs was also found for Groups 3 & 4 and although the value of 0.26 is very small it shows pre-test to post-test improvement in groups using CALL software. In Groups 1 & 2, better performance of CALL groups was also found for subcategory online course (1.07) and CMC + web + course management system (0.87). However, these results are based upon only one effect size each, so additional research is needed in these categories to provide solid evidence.

Examination of effect sizes for other technologies reveals that CALL groups seemed to have performed weaker than non-CALL groups when CMC + course management system and CMC technologies were used. In the case of the former, 12 effect sizes that contribute to the mean effect size of −0.42 come from one study, Chenoweth *et al.* (2006). Eleven of these effect sizes are negative with the lowest values of −0.85, −0.95 and −1.12, while the twelfth effect size is almost zero (0.002).

Table 4 *Effect sizes, standard deviations, and confidence intervals for technology used in studies*

| | Groups 1 & 2 | | | | | Groups 3 & 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 95% Confidence Interval | | | | | 95% Confidence Interval | |
| Technology | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper |
| 1. CALL program | 31 | **0.46** | 0.87 | 0.14 | 0.78 | 4 | **0.26** | 0.32 | −0.25 | 0.77 |
| 2. computer application | 0 | | | | | 3 | **0.92** | 0.15 | 0.55 | 1.29 |
| 3. CMC | 1 | **−0.82** | | | | 3 | **0.03** | 0.59 | −1.44 | 1.49 |
| 4. web | 1 | **0.33** | | | | 5 | **0.74** | 0.27 | 0.41 | 1.08 |
| 5. course management system | 2 | **0.16** | 0.01 | 0.03 | 0.29 | 0 | | | | |
| 6. online course | 1 | **1.07** | | | | 0 | | | | |
| 7. CMC + web + course management system | 1 | **0.87** | | | | 0 | | | | |
| 8. CMC + course management system | 12 | **−0.42** | 0.37 | −0.66 | −0.19 | 0 | | | | |
| 9. computer application + CMC + web + course management system | 0 | | | | | 1 | **0.54** | | | |

Therefore, it is not surprising that the average value is also negative. Similarly, there is only one study, De La Fuente (2003), that gave the effect size value of −0.82 for the CMC subgroup. For Groups 3 and 4, the *ES* value of 0.03 for CMC technology is very close to zero with a very large confidence interval from −1.44 to 1.49. This interval shows that there are similar chances that CALL groups may or may not show improvement from pre- to post-tests after working on CMC tasks. Again, for these categories more comparison CMC studies should be conducted to shed additional light on the issue.

*4.2.2 Degree of integration.* To explore the effects of differing degrees of technology integration in the courses, we coded the studies in the following four subcategories for technology integration:

1) Stand alone technology-based course (all language instruction was provided online in the course).
2) CALL as an add-on component (CALL aspect was administered in addition to instruction in a traditional language course and lasted for only part of the complete course. This component could be more or less integrated into the course).
3) CALL as experiment (CALL aspect was not integrated into language instruction and the duration was very short).
4) Blended learning (language instruction was provided both through CALL and non-CALL methods for the duration of the whole course. The CALL aspect was completely integrated into instruction).

These subcategories were very difficult to define and this was the variable coders most often disagreed on. In cases of disagreement, the coders discussed their choices until the agreement was made.

As Table 5 shows, for Groups 1 & 2, the first three subcategories of integration have effect sizes which are medium or close to medium. The only group with the negative effect size value is blended learning. The close inspection of effect sizes in this group revealed that 12 of them come from Chenoweth *et al.* (2006). These 12 effect sizes bring the average value of the group of 18 effect sizes down (as already explained above). In addition to Chenoweth *et al.* (2006), there is one more negative effect size value in that group (Adair-Hauck *et al.*, 2000). On the other hand, when examining the value for the blended learning subcategory for Groups 3 & 4, it can be seen than the effect size is small to medium but positive throughout. These CALL groups showed positive improvements from the beginning to the end of the study. These contradicting results, for which no explanation is evident, call for more research in the area of blended learning because of the growing importance of blended learning in language classes.

*4.2.3 Length of instructional treatment.* The third aspect of instructional condition investigated was the length of instructional treatment, which was divided into five subcategories based on the number of hours it lasted:

1) Less than 2 hours
2) 2 hours to 4.9 hours

Table 5 *Effect sizes, standard deviations, and confidence intervals for integration of technology*

| | Groups 1 & 2 | | | | | Groups 3 & 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 95% Confidence Interval | | | | | 95% Confidence Interval | |
| Integration of technology | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper |
| 1. Stand alone technology-based course | 3 | **0.45** | 0.54 | −0.91 | 1.80 | 1 | **−0.01** | | | |
| 2. CALL as an add on component | 10 | **0.52** | 0.90 | −0.12 | 1.17 | 5 | **0.37** | 0.63 | −0.41 | 1.16 |
| 3. CALL as experiment | 18 | **0.45** | 0.98 | −0.03 | 0.94 | 0 | | | | |
| 4. Blended learning | 18 | **−0.21** | 0.48 | −0.44 | 0.03 | 10 | **0.63** | 0.33 | 0.39 | 0.86 |

3) 5 to 9.9 hours
4) 10 hours to 16.9 hours
5) More than 17 hours
6) Does not say (length of treatment was not reported).

Table 6 shows that for Groups 1 & 2, medium effect size values were found for less than 2, 2 to 4.9, and more than 17 hours. However, negative values were found for 5 to 9.9 and 10 to 16.9 hours. Again, Chenoweth *et al.*'s (2006) study contributes all 12 effect sizes to the 5 to 9.9 hours subgroup while in the 10 to 16.9 hours subgroup all effect sizes come from different studies, two of which are negative. When we examine the values examine the values for the 10 to 16.9 hours subgroup for Groups 3 & 4, the effect size of 0.77 is around medium, indicating improvements for these CALL groups. Overall, it appears that CALL groups perform better than non-CALL groups and improve over the course of the study no matter how long the treatment lasts.

### *4.3 Learner characteristics*

Learner characteristics of proficiency level and native language were coded to assess any patterns of results that might be explained by these variables.

*4.3.1 Proficiency level.* The following six proficiency levels were used in this study, based on the information from primary studies:

1) beginner
2) intermediate
3) advanced
4) beginner and intermediate
5) intermediate and advanced
6) varied (all three levels were included)

Table 7 contains the results for these subcategories. For Groups 1 & 2, the highest mean effect size is 0.95 for studies with intermediate and advanced learners (3 studies) followed by beginner and intermediate level (3 studies), intermediate (9 studies), and then varied levels (9 studies). Studies examining beginner students have the lowest effect size of −0.01. This is also the largest subgroup of effect sizes, with 23 of them. Nine of those effect sizes come from Chenoweth *et al.* (2006), 3 from Bowles (2004), and 2 from Echavez-Solano (2003) while the other 9 come from 9 different studies. Examination of effect size values for Groups 3 & 4 shows their increase from beginner to advanced level indicating higher levels of improvement in more proficient learners. This result may suggest that the participants' proficiency level makes a difference in study outcomes but needs to be interpreted with caution since there was only one study of advanced learners while there were six and five studies of intermediate and beginner students respectively, in Groups 3 & 4.

*4.3.2 Native language.* The participants' native language had ten subcategories: Spanish, English, Chinese, Japanese, Korean, Arabic, Afrikaans, Danish, and varied

Table 6 *Effect sizes, standard deviations, and confidence intervals for length of treatment*

| | Groups 1 & 2 | | | | | Groups 3 & 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 95% Confidence Interval | | | | | 95% Confidence Interval | |
| Length of treatment | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper |
| 1. Less than 2 hours | 16 | **0.46** | 1.04 | −0.09 | 1.01 | 1 | **0.75** | | | |
| 2. 2 hours to 4.9 hours | 4 | **0.47** | 0.17 | 0.20 | 0.74 | 0 | | | | |
| 3. 5 to 9.9 hours | 12 | **−0.42** | 0.37 | −0.66 | −0.19 | 1 | **0.11** | | | |
| 4. 10 hours to 16.9 hours | 3 | **−0.15** | 0.35 | −1.04 | 0.73 | 6 | **0.77** | 0.24 | 0.52 | 1.01 |
| 5. More than 17 hours | 11 | **0.52** | 0.85 | −0.05 | 1.09 | 7 | **0.30** | 0.56 | −0.21 | 0.82 |
| *Does not say | 3 | | | | | | | | | |

Table 7 *Effect sizes, standard deviations, and confidence intervals for participants' proficiency level*

| | Groups 1 & 2 | | | | | Groups 3 & 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 95% Confidence Interval | | | | | 95% Confidence Interval | |
| Language level | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper |
| beginner | 23 | **−0.01** | 0.72 | −0.33 | 0.30 | 5 | **0.55** | 0.28 | 0.20 | 0.90 |
| intermediate | 9 | **0.37** | 1.01 | −0.41 | 1.14 | 6 | **0.67** | 0.36 | 0.30 | 1.05 |
| advanced | 0 | | | | | 1 | **1.04** | | | |
| beginner and intermediate | 3 | **0.65** | 0.57 | −0.76 | 2.05 | 0 | | | | |
| intermediate and advanced | 3 | **0.95** | 2.06 | −4.17 | 6.07 | 0 | | | | |
| varied (3 levels) | 9 | **0.35** | 0.30 | 0.12 | 0.58 | 2 | **0.08** | 0.14 | −1.15 | 1.31 |
| *Not reported | 2 | | | | | 2 | | | | |

(participants speaking different native languages). The biggest subgroup represents speakers of English learning a foreign language with 9 + 2 effect sizes, followed by Spanish 6 + 2 (see Table 8). However, in 21 + 2 effect sizes primary researchers did not include information about participants' native language so this important piece of information is lost. Table 8 shows that native language does not seem to make a difference because all effect sizes are positive. In sum, CALL groups have outperformed non-CALL groups and improved over the course of the study no matter which native language participants spoke.

### *4.4 Conditions of the research design*

Four conditions of the research design were examined to assess their potential effects on the outcomes of the studies: the setting in which primary research was conducted, language taught, number of participants, and participants' assignment into groups.

*4.4.1 Setting*. The first condition of the research design was the research setting. The following five settings were examined:

1) primary
2) secondary
3) college
4) private language school
5) adult literacy setting

As Table 9 shows, effect sizes for all settings are positive (except for the secondary setting in Groups 3 & 4), indicating that in most educational settings, CALL groups tended to do better than non-CALL groups. The table also shows that the huge majority of studies were conducted in a higher education setting (college) and that primary, secondary, and other settings appear to be underrepresented in CALL research. The college setting also has the majority of studies in Groups 3 & 4 and the effect size of 0.70 with an all positive confidence interval which shows medium improvements for CALL groups.

*4.4.2 Language taught*. There are seven subcategories for language taught in the primary study: ESL, EFL, Spanish, French, German, and Japanese. We coded for English as a second or foreign language based on what primary researchers reported. As can be seen from Table 10, English represents the most commonly taught language followed by Spanish, French, and German. On the other hand, Japanese and Chinese CALL comparison studies were very rare with only one effect size in each group. Very similar effect size values for ESL (0.50) and EFL (0.56) in Groups 1 & 2 and EFL (0.55) in Groups 3 & 4 show positive outcomes for CALL groups learning English. Overall, average effect sizes for Spanish and French are much lower in Groups 1 & 2 but not in Groups 3 & 4. Out of 15 effect sizes for Spanish, 12 come from 4 studies (Chenoweth *et al*., 2006; Bowles, 2004; Aust *et al*., 1993; Echavez-Solano, 2003) so the source of the data needs to be taken into account when considering these results in terms of the overall effectiveness of computer

Table 8 *Effect sizes, standard deviations, and confidence intervals for participants' native language*

| | Groups 1 & 2 | | | | | Groups 3 & 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 95% Confidence Interval | | | | | 95% Confidence Interval | |
| Native language | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper |
| Spanish | 6 | **0.33** | 0.38 | −0.07 | 0.72 | 2 | **0.44** | 0.41 | −3.23 | 4.10 |
| English | 9 | **0.28** | 0.92 | −0.42 | 1.00 | 2 | **0.64** | 0.11 | −0.35 | 1.64 |
| Chinese | 2 | **0.28** | 0.07 | −0.34 | 0.91 | 2 | **0.26** | 1.23 | −10.75 | 11.28 |
| German | 0 | | | | | 0 | | | | |
| Japanese | 3 | **0.40** | 0.12 | 0.09 | 0.71 | 1 | **0.96** | | | |
| Korean | 1 | **0.48** | | | | 0 | | | | |
| Arabic | 1 | **0.81** | | | | 1 | **0.54** | | | |
| Afrikaans | 1 | **0.87** | | | | 0 | | | | |
| Danish | 1 | **1.07** | | | | 0 | | | | |
| varied | 4 | **0.85** | 1.33 | 1.27 | 2.97 | 6 | **0.36** | 0.33 | 0.01 | 0.70 |
| Not reported | 21 | | | | | 2 | | | | |

Table 9 *Effect sizes, standard deviations, and confidence intervals for research setting*

| | Groups 1 & 2 | | | | | Groups 3 & 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 95% Confidence Interval | | | | | 95% Confidence Interval | |
| Setting | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper |
| 1. primary | 6 | **0.37** | 0.38 | −0.03 | 0.77 | 1 | **0.15** | | | |
| 2. secondary | 4 | **0.14** | 0.52 | −0.69 | 0.97 | 3 | **−0.14** | 0.41 | −1.16 | 0.87 |
| 3. college | 37 | **0.21** | 0.94 | −0.10 | 0.52 | 12 | **0.70** | 0.28 | 0.52 | 0.88 |
| 4. private language school | 1 | **0.23** | | | | 0 | | | | |
| 5. adult literacy setting | 1 | **0.34** | | | | 0 | | | | |

Table 10 *Effect sizes, standard deviations, and confidence intervals for language taught in the study*

| | Groups 1 & 2 | | | | | Groups 3 & 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 95% Confidence Interval | | | | | 95% Confidence Interval | |
| Language taught | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper |
| ESL | 15 | **0.50** | 0.72 | 0.10 | 0.90 | 6 | **0.29** | 0.30 | −0.02 | 0.60 |
| EFL | 5 | **0.56** | 0.36 | 0.10 | 1.01 | 4 | **0.55** | 0.80 | −0.70 | 1.81 |
| Spanish | 15 | **−0.06** | 0.83 | −0.52 | 0.40 | 1 | **0.57** | | | |
| French | 8 | **−0.30** | 0.30 | −0.55 | −0.05 | 2 | **0.80** | 0.11 | −0.21 | 1.82 |
| German | 4 | **0.03** | 0.22 | −0.31 | 0.38 | 3 | **0.67** | 0.32 | −0.11 | 1.46 |
| Chinese | 1 | **3.32** | | | | 0 | | | | |
| Japanese | 1 | **0.68** | | | | 0 | | | | |

technology for teaching Spanish. Similar conclusions can be drawn in the case of French teaching because 8 effect sizes come from two studies–Chenoweth *et al*. (2006) seven and Adair-Hauck *et al*. (1999) one effect size.

*4.4.3 Number of participants*. The third condition of the research design was the number of participants in the study. The total number of participants in Groups 1 & 2 is the number of participants in both CALL (experimental) and non-CALL (control) group. Since in Groups 3 & 4 we looked only at the CALL (experimental) group, that is the total number of participants. There are six subcategories for number of participants:

1) less than 20
2) 21–29
3) 30–39
4) 40–59
5) 60–99
6) 100 and more

As Table 11 shows, the effect size values for studies with less than 20 participants, between 40 and 59, 60 and 99, and more than 100 are very similar. These values are small to medium and positive. This may indicate that the number of participants does not make a difference in the language learning outcomes for Groups 1 & 2. The number of effect sizes is not equally distributed for subgroups in 3 & 4, with the category of less than 20 having 11 effect sizes. In any case, all effect sizes are positive and range from very small (0.18) to very large (1.13).

*4.4.4 Method of assignment*. The final condition of the research design was the participants' assignment into groups. Methods of assignment were the following:

1) Random (participants were randomly assigned to conditions).
2) Non-random (participants were non-randomly assigned to conditions, e.g., they were kept in intact classes).
3) Non-random after pairing/matching (participants were first paired based on other criteria, e.g., age, proficiency level, and then non-randomly assigned to conditions).

These three methods of assignment appear on the coding form (see Appendix D online, Part 2, item 12). However, one source, Troia (2004), had some groups assigned randomly and some non-randomly after pairing. Therefore, Table 12 includes the fourth method of assignment (random and non-random after pairing/matching).

Table 12 shows that for both Groups 1 & 2 and Groups 3 & 4, randomization of participants brought the highest mean effect sizes of 0.51 and 0.88 respectively. The magnitude of these effect sizes is medium. It is interesting to note that random assignment does not represent the most common type of assignment since there are 27 effect sizes with non-random assignment in Groups 1 & 2 and 11 in Groups 3 & 4. When assigned non-randomly, CALL groups performed almost the same as

Table 11 *Effect sizes, standard deviations, and confidence intervals for number of participants*

| | Groups 1 & 2 | | | | | Groups 3 & 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 95% Confidence Interval | | | | | 95% Confidence Interval | |
| Number of participants | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper |
| less than 20 | 12 | **0.39** | 1.07 | −0.29 | 1.06 | 11 | **0.56** | 0.46 | 0.25 | 0.87 |
| 21–29 | 7 | **−0.33** | 0.68 | −0.96 | 0.30 | 1 | **0.18** | | | |
| 30–39 | 9 | **0.15** | 0.87 | −0.52 | 0.82 | 0 | | | | |
| 40–59 | 10 | **0.32** | 0.91 | −0.33 | 0.97 | 1 | **1.13** | | | |
| 60–99 | 8 | **0.36** | 0.42 | 0.01 | 0.71 | 3 | **0.22** | 0.29 | −0.48 | 0.94 |
| 100 and more | 3 | **0.39** | 0.45 | −0.72 | 1.51 | 0 | | | | |

Table 12 *Effect sizes, standard deviations, and confidence intervals for method of assignment*

| | Groups 1 & 2 | | | | | Groups 3 & 4 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 95% Confidence Interval | | | | | 95% Confidence Interval | |
| Method of assignment to conditions | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper | K | Mean $E\bar{S}$ | Standard Deviation | Lower | Upper |
| Random | 17 | **0.51** | 1.10 | −0.05 | 1.07 | 2 | **0.88** | 0.22 | −1.11 | 2.87 |
| Non-random | 27 | **0.06** | 0.67 | −0.20 | 0.33 | 11 | **0.48** | 0.49 | 0.15 | 0.81 |
| Non-random after pairing/matching | 4 | **0.18** | 0.22 | −0.17 | 0.53 | 2 | **0.47** | 0.40 | −3.17 | 4.10 |
| Random and non-random after pairing/matching | 1 | **−0.02** | | | | 1 | **0.15** | | | |

non-CALL groups judging by the effect size of 0.06. The improvement of CALL groups assigned non-randomly is small – 0.48 overall – and the all-positive confidence interval shows that effect size gains are estimated to range (with 95% probability) from 0.15 to 0.81. In sum, it appears that random assignment of subjects in CALL studies contributes to larger effect sizes overall.

## 5 Discussion and conclusion

This meta-analysis provides an empirically-based answer to the question of whether pedagogy supported by computer technology can be effective in promoting second/foreign language development relative to pedagogy conducted without technology. Both of these CALL and non-CALL conditions can be realized in many different ways and other methods are needed to investigate the detail of such variations (e.g., Abraham, 2008; Taylor, 2006). Our results showed that across the various conditions of technology use, second/foreign language instruction supported by computer technology was at least as effective as instruction without technology.

When comparisons between CALL and non-CALL groups were made in rigorous research designs (such as those included in Group 1), the CALL groups performed better than the non-CALL groups, as indicated by a small, but positive and statistically significant weighted mean effect size of 0.257. The finding of the strongest effects in this research design may be analogous to the findings of Li (2010), who found the strongest effects in studies of L2 error correction in laboratory settings where the learning conditions could be most strictly controlled relative to classroom conditions. Nevertheless, learning in real classroom conditions is important to study in order to learn the effects of real CALL use by real classroom learners whose purpose is to learn language. From an educational policy perspective, it is important to learn how new innovations in teaching affect learning.

In designs allowing only for calculation of effect sizes based on improvement from pre-tests to post-tests, CALL groups were also found to improve, with effect sizes that were small, but positive and statistically significant. These results corroborate findings from Zhao (2003), Kulik (2003), Liu *et al.* (2002), and Felix (2005b) while including a longer span of time and quantitative findings based on stringent inclusion criteria, which are replicable. They also provide a finding concerning second language learning, which poses learning and teaching challenges different from those of other subjects.

Our exploration of the effects of instructional conditions, characteristics of participants, and conditions of the research design did not provide as reliable results because of the small number of effect sizes (in some cases only one) used as the basis for conclusions. The only clear finding was the difference in studies by research design. As discussed above, research designs employing random placement of subjects into conditions found more marked positive differences for the CALL condition.

One area worthy of further exploration is the effects of proficiency level. In the studies we included, the advanced and intermediate learners did better in the CALL conditions than did beginner learners using CALL, as demonstrated on post-tests. Unfortunately, this trend is only slightly evident because of the lack of studies of advanced learners in Groups 1 & 2, while Groups 3 & 4 had only one study of

advanced learners. This is an area worthy of further investigation in view of the potential implications for curricular decision-making.

The analyses breaking down the studies by instructional conditions, learner characteristics, and research conditions also confirm more precisely the observations made by previous research syntheses. Our finding that higher education tends to be the most frequently researched setting echoes the observations by Zhao (2003), Liu *et al.*, (2002), and Felix (2005a, b), as well as those of the Department of Education report, which underscores the need for research on technology effectiveness in K-12 in the United States. We could add that there is a need for more research that differentiates learners by proficiency level to better understand any potential differences. The breakdown of studies also revealed that ESL/EFL and Spanish represent the most frequently researched languages as previously noted by Zhao (2003) and Felix (2005a). In view of the growth in teaching less commonly taught languages such as Chinese and Arabic in the United States, research is needed in these areas.

Finally, examination of the studies initially found in the search supported the observation made by Zhao (2003) and Felix (2005a) about the lack of methodological rigor in many studies which were excluded from the analysis. In order to include as many primary studies as possible in the future meta-analyses, we strongly suggest randomly assigning participants and when this is not feasible, verifying the comparability of groups with a pre-test at the outset of the study. Moreover, the inclusion criteria and categories compared in this meta-analysis should suggest factors to be considered in designing and reporting primary research.

Future research investigating the effects of technology-supported learning can benefit from such a meta-analysis because it gives a concrete picture of how an individual study can be used to contribute to overall knowledge in the discipline. With respect to CALL, we chose 1970 as a beginning point in order to include all the research that was available on this issue. One finding was that the use of strict criteria for inclusion resulted in the first useable study being conducted in 1984, an unpublished doctoral dissertation at The University of Nebraska – Lincoln, even though other earlier studies had appeared. The useable studies appeared over a period of more than 20 years (1984–2006) and included 37 comparison studies and 144 effect sizes. Given that we have made every effort to provide detailed description of the methodology in this study, we feel future meta-analyses in this area could replicate our findings while expanding the pool of primary studies to include those from January 2007 to the present moment. Additionally, future work could expand on the number of publications searched manually, which was not possible to do in this study.

The present study suggests that a number of questions remain open for investigation, but this summary of results from the beginning of CALL indicate that pedagogical options that computer technology offers language learners are worthy of further investigation. Perhaps more important for educational decision-making today, the overall results did not indicate that CALL was inferior to classroom conditions.

## Supplementary materials

For supplementary material referred to in this article, please visit http://dx.doi.org/10.1017/S0958344013000013

## References


Abraham, L. B. (2008) Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning*, **21**: 199–226.

Adair-Hauck, B., Willingham-McLain, L. and Youngs, B. E. (2000) Evaluating the integration of technology and second language learning. *CALICO Journal*, **17**(2): 296–306.

Al-Jarf, R. S. (2002) Effect of online learning on struggling ESL college writers. *National Educational Computing Conference*, San Antonio, Texas: National Educational Computing Conference Proceedings, 23.

Aust, R., Kelley, M. J. and Roby, W. (1993) The use of hyper-reference and conventional dictionaries. *Educational Technology, Research and Development*, **41**(4): 63–73.

Bowles, M. A. (2004) L2 glossing: To CALL or not to CALL. *Hispania*, **87**(3): 541–552.

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

Chapelle, C. A. (2003) *English language learning and technology: Lectures on applied linguistics in the age of information and communication technology*. Amsterdam: John Benjamins Publishing.

Chapelle, C. A. (2007) Challenges in evaluation of innovation: Observations from technology research. *Innovation in Language Learning and Teaching*, **1**(1): 30–45.

Chenoweth, N. A., Ushida, E. and Murday, K. (2006) Student learning in hybrid French and Spanish courses: An overview of language online. *CALICO Journal*, **24**(1): 115–145.

Cooper, H. (1998) *Synthesizing research: A guide for literature reviews*. Thousand Oaks, CA: Sage.

Cohen, J. (1988) *Statistical power analysis for behavioral sciences*. Hillsdale, NJ: Erlbaum.

De la Fuente, M. J. (2003) Is SLA Interactionist theory relevant to CALL? A study on the effects of computer-mediated interaction in L2 vocabulary acquisition. *Computer Assisted Language Learning*, **16**(1): 47–81.

Dunkel, P. (1991) The effectiveness research on computer-assisted instruction and computer-assisted language learning. In: Dunkel, P. (ed.), *Computer-assisted language learning and testing: Research issues and practice*. New York: Newbury House, 5–36.

Echavez-Solano, N. (2003) A comparison of student outcomes and attitudes in technology-enhanced vs. traditional second-semester Spanish language courses. Unpublished doctoral dissertation, The University of Minnesota.

Felix, U. (2005a) What do meta-analyses tell us about CALL effectiveness? *ReCALL*, **17**(2): 269–288.

Felix, U. (2005b) Analyzing Recent CALL Effectiveness Research – Towards a Common Agenda. *Computer Assisted Language Learning*, **18**(1–2): 1–32.

Jeon, E. H. and Kaya, T. (2006) Effects of L2 instruction on interlanguage pragmatic development. In: Norris, J. and Ortega, L. (eds.), *Synthesizing research on language learning and teaching*. Philadelphia PA: John Benjamins, 165–211.

Kulik, J. A. (2003) *Effects of using instructional technology in colleges and universities: What controlled evaluation studies say*. Arlington: SRI International.

Kulik, J. A. and Kulik, C. L. C. (1987) Review of recent literature on computer-based instruction. *Contemporary Education Review*, **12**: 222–230.

Kulik, J. A. and Kulik, C. L. C. (1991) Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior*, **7**(1–2): 75–94.

Li, S. (2010) The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, **60**(2): 309–365.

Liu, M., Moore, Z., Graham, L. and Lee, S. (2002) A look at the research in computer-based technology use in second language learning: A review of literature from 1990–2000. *Journal of Research on Technology in Education*, **34**(3): 250–273.

Lipsey, M. W. and Wilson, D. B. (2001) *Practical meta-analysis*. Thousand Oaks, CA: Sage.
Nagata, N. (1996) Computer vs. workbook instruction in second language acquisition. *CALICO Journal*, **14**(1): 53–75.
Norris, J. M. and Ortega, L. (2000) Effectiveness of L2 instruction: A research synthesis and a quantitative meta-analysis. *Language Learning*, **50**: 417–528.
Norris, J. M. and Ortega, L. (2006) *Synthesizing research on language learning and teaching*. Philadelphia, PA: John Benjamins.
Odenthal, J. M. (1992) The effect of a computer-based writing program on the attitudes and performance of students acquiring English as a Second Language. Unpublished doctoral dissertation, Claremont Graduate School, San Diego State University, Claremont, California.
Oswald, F. L. and Plonsky, L. (2010) Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, **30**: 85–110.
Payne, J. S. and Whitney, P. J. (2002) Developing L2 oral proficiency through synchronous CMC: output, working memory, and interlanguage development. *CALICO Journal*, **20**(1): 7–32.
Pederson, K. M. (1987) Research on CALL. In: Smith, W. F. (ed.), *Modern media in foreign language education: Theory and implementation*. Lincolnwood, IL: National Textbook Company, 99–132.
Rawdon, T., Sharp, R. L., Shelley, M. and Thomas, J. R. (2012) Meta-analysis of the placebo effect in nutritional supplement studies of muscular performance. *Kinesiology Review*, **1**(2): 137–148.
Shelley, M., Rawdon, T., Sharp, R., Thomas, J., and Schalinske, K. (2007) Meta-analysis in the human sciences: Are placebo effects real? In: American Statistical Association, *2006 proceedings of the American Statistical Association, section on survey research methods* [CD-ROM]. Alexandria, VA: American Statistical Association.
Taylor, A. M. (2006) The effects of CALL versus traditional L1 glosses on L2 reading comprehension. *CALICO Journal*, **23**: 309–318.
Tozcy, A. and Coady, J. (2004) Successful learning of frequent vocabulary through CALL also benefits reading comprehension and speed. *Computer Assisted Language Learning*, **17**(5): 473–495.
Troia, G. A. (2004) Migrant students with limited English proficiency: Can *Fast ForWord Language* make a difference in their language skills and academic achievement? *Remedial and Special Education*, **25**(6): 353–366.
US Department of Education, Office of Planning, Evaluation, and Policy Development (2009) *Evaluationof evidence-based practices in online learning: A meta-analysis and review of online learning studies*. Washington, DC. www.ed.gov/about/offices/list/opepd/ppss/reports.html
Zhao, Y. (2003) Recent developments in technology and language learning: a literature review and meta-analysis. *CALICO Journal*, **21**(1): 7–27.

## Appendix E

### *Studies included in the meta-analysis*

Adair-Hauck, B., Willingham-McLain, L., & Youngs, B. E. (2000). Evaluating the integration of technology and second language learning. *CALICO Journal,* **17**(2): 296–306.
Al-Jarf, R. S. (2002). Effect of online learning on struggling ESL college writers. *National Educational Computing Conference* San Antonio, Texas: National Educational Computing Conference Proceedings, 23.

Al-Juhani, S. O. (1991). The effectiveness of computer-assisted instruction in teaching English as a Foreign Language in Saudi secondary school. Unpublished doctoral dissertation, University of Denver, Colorado.

Aust, R., et al. (1993). The use of hyper-reference and conventional dictionaries. *Educational Technology, Research and Development,* **41**(4): 63–73.

Bowles, M. A. (2004). L2 glossing: To CALL or not to CALL. *Hispania,* **87**(3): 541–552.

Cahill, D., & Catanzaro, D. (1997). Teaching first-year Spanish on-line. *CALICO Journal,* **14**(2-4): 97–114.

Cartez-Enriquez, N., Rodriguez, M.I.S., & Quintana, L.R. (2004). Electronic texts or learning through textbooks: an experimental study. *ReCALL,* **16**(2): 539–554.

Chenoweth, N. A., Ushida, E., & Murday, K. (2006). Student learning in hybrid French and Spanish courses: An overview of language online. *CALICO Journal,* **24**(1): 115–145.

Chiappone, L. L. (2004). A comparative study of English language learners reading story-books in traditional print and digital formats. Unpublished doctoral dissertation, Dissertation Abstracts International, 200505115.

Chuo, T. I. (2004). The effect of the webquest writing instruction on EFL learners' writing performance, writing apprehension, and perception. Unpublished doctoral dissertation, School of Education, La Sierra University, Riverside, California.

Coniam, D., & Wong, R. (2004). Internet Relay Chat as a tool in the autonomous development of ESL learners' English language ability: an exploratory study. *System,* **32**(3): 321–335.

De la Fuente, M. J. (2003). Is SLA Interactionist theory relevant to CALL? A study on the effects of computer-mediated interaction in L2 vocabulary acquisition. *Computer Assisted Language Learning,* **16**(1): 47-81.

Dreyer, C., & Nel, C. (2003). Teaching reading strategies and reading comprehension within a technology-enhanced learning environment. *System,* **31**(3): 349–365.

Echavez-Solano, N. (2003). A comparison of student outcomes and attitudes in technology-enhanced vs. traditional second-semester Spanish language courses. Unpublished doctoral dissertation, The University of Minnesota.

Felix, U., & Lawson, M. (1996). Developing German writing skills by way of Timbuktu: A pilot study comparing computer-based and conventional teaching. *ReCALL,* **8**(1): 12–19.

Ghaleb, M. L. N. (1993). Computer networking in a university freshman ESL writing class: A descriptive study of the quantity and quality of writing in networking and traditional writing classes. Unpublished doctoral dissertation, The University of Texas at Austin, Austin, Texas.

Green, A., & Youngs, B. E. (2001). Using the web in elementary French and German courses: Quantitative and qualitative study results. *CALICO Journal,* **19**(1): 89–123.

Hong, W. (1997). Multimedia computer-assisted reading in business Chinese. *Foreign Language Annals,* **30**(3): 335–344.

Kang, S.-H. (1992). Computers and context-embedded language learning. *CAELL Journal,* **3**(3): 15–20.

Kim, M. J. (1993). *The MacMagic program and its effects on "English as a Second Language" students. An evaluation study*. Marin Community Foundation, San Rafael, CA., ED372630.

King, M. H. (1985). The impact of computer assisted instruction on the acquisition of English as a second language. Unpublished doctoral dissertation, School of Education, United States International University, San Diego, California.

Leffa, V. J. (1992). Making foreign language texts comprehensible for beginners: an experiment with an electronic glossary. *System,* **20**(1): 63–73.

Liou, H. (1997). The impact of www texts on EFL learning. *Computer Assisted Language Learning,* **10**(5): 455–478.

Liou, H., Wang, S.H., & Hung-Yeh. Y. (1992). Can grammatical CALL help EFL writing instruction? *CALICO*, **10**(1): 23–44.

Mellgren, M. P. (1984). The effects of supplemental computer instruction on achievement in Spanish cognitive style, field-dependence, foreign language. Unpublished doctoral dissertation, The University of Nebraska - Lincoln, Lincoln, Nebraska.

Nagata, N. (1996). Computer vs. workbook instruction in second language acquisition. *CALICO Journal,* **14**(1): 53–75.

Niwa, Y., & Aoi, K. (1990). A preliminary investigation into the efficiency of CAI. *IALL Journal of Language Learning Technologies,* **23**(3): 9–20.

Odenthal, J. M. (1992). The effect of a computer-based writing program on the attitudes and performance of students acquiring English as a Second Language. Unpublished doctoral dissertation, Claremont Graduate School, San Diego State University, Claremont, California.

Payne, J. S., & Whitney, P. J. (2002). Developing L2 oral proficiency through synchronous CMC: output, working memory, and interlanguage development. *CALICO Journal,* **20**(1): 7–32.

Petersen, M. J. (1990). An evaluation of VOXBOX, a computer-based voice- interactive language learning system for teaching English as a Second Language. Unpublished doctoral dissertation, School of Education, United States International University, San Diego, California.

Spelman, M. D. (2002). GLOBECORP: simulation versus tradition. *Simulation & Gaming,* **33**(3): 376-394.

Stenson, N., Downing, B., Smith, J., & Smith, K. (1992). The effectiveness of computer-assisted pronunciation training. *CALICO*, **9**(4): 5–19.

Stoehr, L. E. (2000). The effects of built-in comprehension aids in a CALL Program on student-readers' understanding of a foreign language literary text. Unpublished doctoral dissertation, The University of Texas at Austin, Austin, Texas.

Taniguchi, M., & Abberton, E. (1999). Effect of interactive visual feedback on the improvement of English intonation of Japanese EFL learners. *Speech, Hearing and Language,* **11**: 76-89.

Terhune, D. R., & Moore, B. (1991). Computer vs. paper: a preliminary study on vocabulary expansion. *CAELL Journal,* **2**(3): 30–34.

Tozcu, A., & Coady, J. (2004). Successful learning of frequent vocabulary through CALL also benefits reading comprehension and speed. *Computer Assisted Language Learning,* **17**(5): 473–495.

Troia, G. A. (2004). Migrant students with limited English proficiency: Can *Fast ForWord Language* make a difference in their language skills and academic achievement? *Remedial and Special Education,* **25**(6): 353–366.

**Glossary of main statistical terms and research designs**

| Statistic name | Definition |
|---|---|
| ANCOVA | Analysis of covariance, used with a continuous dependent variable and a combination of categorical and continuous predictor variables. |
| Between-subjects design | A type of research design in which the control and experimental group consist of different participants. |
| Control group | The study group that did not receive the treatment. |
| Effect size | A statistic that standardizes study findings and allows for more meaningful comparison across studies. Expresses the difference between groups or over time. |
| Experimental design | A type of research design in which participants are randomly assigned to groups. |
| Experimental group | The study group that receives the treatment. |
| *F*-value | The test result that indicates whether there is a significant difference in outcomes across groups. |
| Mean | A statistic that shows the average value of a variable. |
| *p*-value | A statistic that shows whether a statistical test indicates a statistically significant outcome. |
| Pooled standard deviation | An averaged value of the standard deviations from two or more groups. |
| Q value | A statistic that indicates the dispersion of effect sizes around the mean and whether they follow a normal distribution. Also known as homogeneity. |
| Quasi-experimental design | A type of research design in which participants are not assigned to groups at random; commonly used when groups occur naturally. |
| Scheffé and Bonferroni multiple comparisons | Scheffé multiple comparisons are particularly useful in analysis of variance and in constructing simultaneous confidence bands for regressions. Scheffé's method is a single-step multiple comparison procedure that applies to the set of estimates of all possible contrasts among the factor level means, not just the pairwise differences.<br><br>Bonferroni multiple comparisons are the most conservative method to control the familywise error rate. If it is desired that the significance level for the whole family of n tests should be (at most) $\alpha$, then the Bonferroni approach is to test each of the individual tests at a significance level of $\alpha/n$. |
| Standard deviation | The average distance between individual test scores and the mean test score. Expresses how much, on average, individual results deviate from the mean. |

*Continued*

| Statistic name | Definition |
|---|---|
| Standardized mean difference | The difference between the means of experimental and control groups on a dependent variable following implementation of the treatment, divided by the pooled standard deviation. This is expressed commonly as Cohen's d measure of effect size for independent groups. |
| Standardized mean gain | The difference between the pre-implementation and post-implementation means of the experimental or control group on a dependent variable divided by the pooled standard deviation. This is commonly expressed as Cohen's d measure of effect size for change over time. |
| Tamhane procedure | Tamhane multiple comparisons assume unequal group variances, whereas Scheffé and Bonferroni both assume equal variances. Tamhane's procedure provides conservative pairwise comparisons based on a t-test. |
| *t*-value | A statistic that indicates whether there is a significant difference between groups or a significant change over time; consists of a difference between means divided by the standard error of that difference. |
| Weighted effect size | Adjusts for the number of observations on which each effect size was calculated; more weight is given to larger sample sizes. |
| Within-subjects design | A type of research design in which outcomes for the same participants are measured at least twice over time. |