ReCALL **17** (2): 269–288. © 2005 Cambridge University Press
DOI: 10.1017/S0958344005000923 Printed in the United Kingdom

What do meta-analyses tell us about CALL effectiveness?

USCHI FELIX
Monash University, Melbourne, Australia
(*email: uschi.felix@arts.monash.edu.au*)

Abstract

A great deal of research has been carried out over the past two decades related to the effectiveness of computer-assisted teaching and learning. This large body of work, however, has not produced unequivocal findings, especially in the area of CALL. Our study takes a systematic look at what dedicated meta-research since 1991 might have contributed to this controversial field. Data was gathered from several hundred studies including ca 20,000 subjects. The paper highlights the benefits and limitations associated with this type of research, synthesizes findings and recommends models and strategies for future research in the context of an agreed research agenda. Our synthesis suggests consistent positive findings related to L1 spelling, writing and reading. While there is some support for this related to CALL per se, much work needs to be done before we might reach the levels of confidence enjoyed by some of the findings reported here.

1 Introduction

Why do we still know so little about the efficacy of the technologies into which we have invested much energy, time and money in our teaching and learning endeavours? A question that sounds quite simple on the face of it is, of course, immensely complex when we look at it more closely. How can we hope to arrive at a sensible answer to the often posed question *How effective are technologies in promoting learning?* when the scope of the investigations required to come even close to a valid generic conclusion is beyond most researchers' capacity. We might as well capitulate and say that it is '42' which was writer Douglas Adams' answer to the question of life, the universe and everything.

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

It appears that we have two choices in tackling the question which remains interesting and worth investigating, if only to confirm our instincts that what we are doing with a great deal of effort is worthwhile. On the one hand, we can reduce the scope by focusing on one particular piece of technology being used having one or several effect(s) on a measurable learning process or outcome. On the other hand, we can widen the scope by analyzing results of a large body of research and synthesizing findings related to one or several variable(s) under investigation. In both cases, the question needs to be qualified and matched more closely to the project's context, methods and analyses. Curiously, all of this has already been done over the past two decades, yet general conclusions and claims remain largely controversial.

The 1980s saw great interest in meta-analyses being carried out into CAI in general (mainly in science and mathematics), a field led by Kulik *et al.* (1980, 1987), with positive effects being reported on some learning outcomes, such as higher achievement and time-saving as well as improved attitude to computers and learning. In CALL, meta-research is relatively scarce – we identified a total of thirteen studies (see 'further reading') for the current investigation and Dunkel (1991), in her seminal chapter on CALL effectiveness, reports findings of a narrative review (Hope *et al.*, 1984) and a single meta-analysis (Roblyer *et al.*, 1988). She concludes that the first researchers were unable to draw unequivocal findings (Dunkel, 1991: 16) and questions the findings of the second study (ibid: 17) on the basis that the three investigations involved contained contradictory results and that other important work had not been included.

It appears that on the whole CALL researchers are most interested in small-scale experimental, quasi-experimental and non-experimental studies and evaluations of a specific intervention by way of Information and Communication Technologies (ICT). In the course of our previous work (Felix, 2005), which looked at 52 such studies carried out since 2000, it became clear that we are beginning to see a substantial and valid body of qualitative and quantitative data related to CALL effectiveness (as defined in section 3.1). While concerns expressed by Chapelle and Jamieson (1991) and many others regarding the lack of random assignment of subjects and control for extraneous variables are still present (see Felix (2005) for a detailed discussion), they are no longer sustainable to the same extent. Our sample showed an almost equal distribution between experimental and non-experimental studies. This can be seen as a strength of the research because their combined results advance the field more substantially than a continuous series of highly controlled studies, especially since the latter are predominantly very short-term (sometimes as brief as a 40-minute treatment). It is important to keep in mind Kulik and Kulik's (1991) finding that the treatment feature most directly related to effect size is length of treatment. In their studies, computer-based instruction was especially effective when the duration of treatment was four weeks or less (Kulik & Kulik, 1991: 88). In this context it is important to mention that we also know very little about how long reported improvement is maintained (Felix, 2005).

As predicted by Dunkel (1991), our previous study identified a noticeable move towards the inclusion of learning processes. This is not surprising since teaching methods and technologies have changed (albeit not everywhere) in a way that no longer allows for easy numeric measurement of outcomes. Studies included the investigation of variables such as *learning strategies and styles, motivation, time-on-task, attitude, conversational interaction, autonomy, critical reflection, perceptions, negotiation and*

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

noticing. While these projects have much merit in their own right, it goes without saying that synthesizing results of this body of work as a whole would not yield very coherent conclusions related to generic effectiveness. However, this work provides opportunities for building up a body of clustered research results which might give more detailed insights into the impact of ICT on any particular process or outcome. If we look at the sixteen studies that investigate writing identified in Felix (2005), for instance, nine are concerned specifically with writing quality. A synthesis of those findings could potentially produce more powerful quantitative conclusions than one single study. However, since the studies in question are spread fairly evenly across the spectrum of research designs and use various forms of qualitative and quantitative measures, calculating effect sizes (defined in section 3.3) would appear impossible. But should this, in fact, concern us? What is it that meta-analyses contribute to the field? Can they produce less equivocal results than those reported extensively in Dunkel (1991) and bring us closer to a credible answer?

2 Purpose of the study

Essentially our study aims to throw light on a continuously confusing body of research by analyzing what has been contributed to the question of CALL effectiveness by way of meta-analyses and reviews since 1991. This information will fill a significant gap in the research by examining a substantial body of studies not included in either Dunkel (1991) or Felix (2005). We expect that drawing conclusions in the light of these extensive previous studies will produce a comprehensive picture of what has been investigated and found related to CALL effectiveness since the 1980s. It is hoped that these results will be of use to researchers in planning their studies on the basis of what has already been done in their area of interest and thus strengthen their work. For this purpose, we recommend that the current findings are viewed together with those reported in Dunkel (1991) and Felix (2005). In the following we will:

- define what is meant by *effectiveness, meta-analysis* and *effect size*
- identify strengths and weaknesses of meta-analysis research
- describe the literature search and the criteria for inclusion
- summarize major findings arrived at by means of meta-analyses and reviews since 1991
- single out valid meta-analysis procedures that may serve as models for researchers
- draw conclusions and suggest directions for future research in CALL effectiveness.

3 Definitions

3.1 Definition of effectiveness

The Webster dictionary defines *effective* (from *efficere*) as 'having an effect, producing a result, bringing something to pass'. *Efficacy* (synonymous with effectiveness) is defined as 'the power to produce effects or intended results'. This suggests a strong causal relationship between an intervention, such as the use of a particular item of

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

technology in a learning situation and a discernible change in the learning process, the learning climate or the learning achievement. Definitions given in the other major dictionaries (Oxford, Microsoft Encarta, Collins) support this view.

3.2 Definition of meta-analysis

A meta-analysis is a synthesis of the findings of several experimental studies. It uses statistical techniques for aggregating the results of multiple empirical studies to determine the direction and size of relationships between similar variables across these studies. Important characteristics of meta-analyses are that they:

- use quantitative measures
- do not prejudge research findings in terms of research quality
- seek general conclusions in relation to a common issue.

For more detailed information see Glass *et al*. (1981), Cooper and Hedges (1994), Schafer (1999) and Suri (2000).

3.3 Definition of effect size

The basis of statistical techniques used in meta-analysis is the calculation of an effect size; a standardized measure which indicates the extent to which experimental and control groups differ in the mean values of a dependent variable at the end of a treatment phase. See Figure 1 for an accessible, comprehensive description by Miech *et al*. (1997).

In most meta-analyses the effect size is reported as the *average* or *mean* effect size. Sometimes, however, the *median* effect size is reported. The mean or average is calculated by summing all the effect sizes and dividing by the number of effect sizes summed. The median is a measure of central tendency, representing the point at which exactly half the effect sizes are below that figure and half are above. In order for this to be meaningful, some indication of the range of values is needed. For example if there are five effect sizes 1.12, 0.95, 0.25, 0.05, 0.01, the median is 0.25 and the average 0.476. If the five effect sizes are 0.45, 0.35, 0.25, 0. 15, 0.05 the median is still 0.25 but the average is also 0.25. In the absence of the full range of values it is, therefore, not possible to interpret results accurately (see findings in section 6).

4 Strengths and weaknesses of meta-analysis research

While the generalizability of a single study is often restricted by sampling characteristics such as sample size and population, study setting and timing, a meta-analysis, by including all the quantitative empirical studies relevant to the research question, enhances generalizability. It quantifies the effect of a treatment (i.e. the use of a particular form of technology), and it should be free from the subjectivity introduced by selective sampling (Suri, 2000). Ideally, meta-analyses provide a statistically sound summary review of a large body of research related to a common issue.

However, there are a number of limitations to meta-analysis. Cooper and Hedges (1994) point out that the *post hoc* nature of synthesis tests often creates a conflict when

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

EFFECT SIZES

A common method of reporting results of empirical studies and meta-analyses is to use effect size. In brief, an effect size is a simple way to compare the outcomes among studies with differing numbers of participating students by standardizing the results. For example, an effect size of .5 would mean that, on average, students formerly at the 50th percentile would now achieve at the 69th percentile. An effect size of 1.0 would move the median student to the 84th percentile. Educators generally agree that an effect size of around .3 (a move to the 62nd percentile for the median student) or larger represents a substantial education benefit, especially when we consider that these effect sizes represent the average improvement for a *population* of students, not just one student. However, the merit of a given effect size for any education intervention also depends on what other options may be available and on their relative costs.

Sample Effect Sizes and Their Related Percentile Differences

Effect Size	.00	.10	.20	.30	.40	.50	.75	1.0	1.5	2.0
Percentile	50	54	58	62	66	69	77	84	93	98

Fig. 1. an accessible, comprehensive description by Miech *et al.* (1997).

stating a hypothesis based on the considerable knowledge that a researcher already has of the data to be synthesized, and then using that data to support the hypothesis. Another problem is that meta-analysis favours published over unpublished research and it has been speculated that journals favour studies that report significant positive outcomes (Suri, 2000; Zhao, 2001). This may skew the findings of meta-analyses towards positive results. Also, studies for which the effect size cannot be computed are ignored by researchers who produce meta-analyses (Suri, 2000). Important results, arrived at through qualitative measures, are therefore not included in reporting findings on a particular issue under investigation. The most serious concern, often pointed out by critics of meta-analyses, is the correlational nature of the review evidence. Including studies that use different procedures to test the same hypothesis can lead to confounds which, despite statistical control, result in low confidence levels in the accuracy of the outcomes. Many reviewers fail to "study whether the findings of the research were mediated by characteristics of the person studied, the study context, the nature of the experimental intervention, or the characteristics of the research design" (Glass *et al.*, 1981:13). In summary, meta-analyses are by their very nature prone to overgeneralization and sometimes include results from poorly designed studies.

5 Literature search and criteria for inclusion

The literature search for our study was carried out by conducting a search of the following computerized databases: Proquest, Educational Resources Information Centre (ERIC), Expanded Academic ASAP, CSA Illumina and ProQuest Digital Dissertations. Three groups of keywords were used for the search. The first was for the type of study ("meta-analysis", "research synthesis", "review"); the second defined the technology ("computer assisted", "CAI," "Internet", "multi-media"); and the third the dependent variable ("language learning", "reading", "writing", "spelling", "vocabulary", "grammar",

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

"literacy"). The search was restricted to papers published after 1991 and written in English. The reason for the 1991 cut-off was that the studies reported here include investigations going back as far as 1980 and that Dunkel (1991) gives a comprehensive overview of reported effectiveness in studies not included here (incorporated into the discussion below). Including articles written in other languages was at this stage beyond our capacity. Since the work is ongoing, however, we welcome relevant information. Our searches resulted in more than 150 papers. Most of these were discarded because they were not concerned with effectiveness or were not research reviews. This left thirteen studies which sought to produce an overview of effectiveness research related to some aspect of language learning, including L1 reading and writing. Eight of these were meta-analyses (calculating effect sizes) and five were reviews (reporting findings discursively). We included the latter to see whether these qualitative syntheses contributed support and/or further information to the quantitative results of the former. The L1 studies were included because they investigated much the same variables (i.e. writing quality, number of words) as the L2 studies and their findings would be most relevant to CALL.

6 Findings

6.1 General observations

We were surprised by the glaring lack of meta-research related to CALL *per se*. Our search resulted in only one meta-analysis (Zhao, 2001) and three narrative reviews (Liu, Moore *et al*., 2002; Miech *et al*., 1997; Singhal, 1998). Together with the two studies discussed in Dunkel (1991) this constitutes a relatively small body of data when compared to meta-research carried out in general education. Perhaps one reason for the scarce research of this type is that the CALL profession, in general, is less interested in effect sizes and more concerned with looking at how the use of technologies might affect learning processes. The type of research carried out most recently (see Felix, 2005) supports this view quite strongly. Another reason might be that the time-consuming research designs, data gathering and statistical analyses necessary are beyond the notoriously overworked CALL practitioners who constitute the bulk of researchers in this field.

In terms of settings, it is interesting to note that L2 studies were predominantly carried out in tertiary learning environments while the majority of L1 studies were conducted in primary and secondary schools. This confirms previous observations made in Dunkel (1991) and Felix (2005). It is not surprising, since most CALL research is carried out by practitioners located in universities, drawing on student populations in their own or related courses. Substantial, school-based educational research, by contrast, tends to be conducted by large, sometimes independent, research teams, drawing on large student numbers from a variety of school settings.

One concern that springs to mind when glancing at the summary tables is that there is likely to be some overlap in the reported numbers of studies and students, since some of the meta-studies share the same time period from which they drew their data. Unfortunately, we cannot be precise here because not all primary studies are listed in all secondary ones, especially in the very large investigations. We are therefore reluctant to report an exact overall number of projects and subjects involved in the 13 investigations.

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

It can be estimated, however, that several hundred studies were included and that student numbers exceeded 20,000. Kulik (2003) alone reports 20,000 subjects.

Even though this number is impressive, we cannot be certain that all relevant research studies were included. Successful searches are very largely dependent on the match between keywords entered and those present in relevant research studies. The difficulty of homing in on just the right keywords became obvious in our own searches for this and previous research. Covering all relevant databases is also imperative (this was not uniformly the case in this body of work). The sample also does not include unpublished and current work or papers written in languages other than English.

A very frustrating aspect of our review was that the thirteen projects report variables under investigation in vastly different ways. They range from providing great detail, i.e. every common aspect of writing (*quality, number of words, attitude towards writing, writing conventions, frequency of revision* in Bangert-Drowns (1993)), minimal descriptions of what might be meant by *Reading Management* (Kulik, 2003),[1] to generic statements such as *English* (Christmann & Badgett, 1997; Christmann *et al*., 1997). It goes without saying that the first yields more interesting information than the others. Smaller meta-analyses dealing with one discrete aspect of learning usually give the most detailed information on the variable under investigation. Very large studies dealing with a variety of subjects, by contrast, tend to report these simply as *music, mathematics,* or *English*. We have made every effort to include as much information as could be ascertained in the summary tables but realize that in some cases more detail would have been helpful for meaningful interpretation of results. In the following we will report findings separately for meta-analyses and reviews.

6.2 Meta-analyses

We have divided the discussion of findings into three categories: (1) studies concerned specifically with CALL; (2) studies dealing with general subjects but including findings relevant to L1 learning; and (3) studies concentrating on just one aspect of L1 learning (see Table 1 for summary data).

6.2.1 CALL

Study 5 is the only meta-analysis concerned with CALL *per se* but also sadly the least satisfactory in terms of interpretable outcomes. The only thing the studies chosen for this meta-analysis have in common is the fact that they test for the effectiveness of some aspect of CALL in a wide-ranging, undefined collection (*listening, reading, writing, speaking, culture, vocabulary* and *grammar*). This could explain the very high effect size (1.12) which is very much at variance with all other findings reported in our current paper. If the studies are grouped according to the variable being tested then the result would be considerably different. For instance, two studies deal with *vocabulary*: one has an effect size of 1.11, the other 0.38 giving an average of 0.75 – considerably less than

1. Reading management programs. Reading management programs, such as Accelerated Reader (AR), help students make book selections and then test the students on their understanding of what they have read (Kulik, 2003:vi).

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

Table 1. *Summary data*

Id	Year Published	Period encompassed	Ed Setting	Nº Studies	Nº Students	Variables	Effect size
1	1993	1983-1990	All	33	> 1874*	L1 Writing using Word Processing	Quality: 0.27 Length : 0.52 Conventions: 0.16 Frequency of Revision: 0.18
2#	1997	1984-1995	Secondary	27	3785	L1 Reading L1 English	Reading: 0.262 English: -0.420
3#	1997	1984-1995	Secondary	26	3694	Differentiated effect size over time	Overall: 0.187 Over time decreases from -0.455 to 0.884
4	2000	1982-1999	Primary & Secondary	17	1590	L1 Reading	Overall: 0.13
5	2001	1997-2001	Tertiary	9	419	L2 CALL effectiveness	Overall: 1.12
6	2002	1991-2000	Primary	42	>3605*	L1 Reading	Overall: 0.25
7	2002	1980-2000	Primary	7	240	L1 Spelling	Overall: 0.37
8#	2003	1990-1999	Primary & Secondary	9	>3620*	Integrated learning systems	Median 0.06
				13	>3300*	Writing based reading programs	
						Kindergarten	Average 0.84
						Grade 1	Average 0.40
						Grades 2 & 3	Median 0.25
				6	>10000**	Reading management	Median 0.43
				10	>1325***	Writing – Word Processing	Median 0.30
						Writing – Computer writing prompts	Not given
						Writing – Computer enrichment	Median 0.34

* some studies do not report the number of students ** 3 studies report number of schools (7,240) *** One study reported 16 schools
# includes non-language related variables

the reported overall effect size, although calculating average effect sizes of only two studies is not advisable in general (see section 6.3). It also needs to be pointed out that the overall effect size is greatly boosted by the two studies on *culture* (non-specified and both carried out by the same researchers in the same year) – 2.82 and 1.61. If these are taken out of the calculation then the overall effect size for the remaining seven studies is 0.81. While this would suggest a positive impact of whatever technologies had been used on L2 learning in general, it tells us a lot less than a study which might have looked at all investigations of word learning, including all possible information on technologies and methods used. To be fair to the authors, it is quite difficult to gather this information because primary studies often do not provide it and the study as it stands is still valuable when we can look at the results in the context of a large body of work, in which the question of effectiveness is examined through a variety of approaches.

6.2.2 General subjects

Studies 2, 3 and 8 take a broad look at effectiveness across a number of subjects such as music, science and mathematics, but include a break-down of effect sizes related to individual subjects or learning aspects. We have included here specific findings related to language learning. While these studies are more satisfactory in terms of methodology than **5** above, there are, nevertheless, some issues that need to be pointed out.

- Findings of study 8 are difficult to interpret because the author indiscriminately alternates between reporting average and median effect sizes. He also erroneously reports a median effect size as an average effect size (results for Grade 1 studies on *writing-based reading programs*). In the absence of the full range of values, we could not, unfortunately, re-analyse the data (see discussion on effect sizes in section 3.3) and results have to be viewed with some caution.
- The subject "English" which was part of the investigations in study 2 is so broad that reporting an overall effect size (which incidentally is greatly at variance with all other findings reported in our sample) is not very meaningful generally and thus of little value to us here. Equally, the mean effect size of 0.209 for the 42 effect sizes given (p 285) is not terribly useful, given the wide range of subjects, and presumably variables, covered. Reporting the overall effect size on the more tangible variable such as *reading*, by contrast, appears much more useful and is included in our discussion.
- The authors of study 2 re-analysed the data for a second paper (3) to ascertain whether effect sizes changed over time. This is potentially a very interesting question but highly problematic in terms of interpretation. While the authors here conclude that there is a noticeable decline in effectiveness over time – culminating in a substantial negative effect (-0.45) – these findings need to be looked at more closely. Calculating effect sizes over time is quite problematic as each time period does not include the same study population, settings and approaches. Also for most years there are fewer than five studies included – in fact for three years (1984, 1988, 1993) there is only one study. The 1985 figure includes just two language-related studies, one English and one Reading, with an average effect size of 0.30. If we look only at the language-related subjects, the effect sizes over time are as shown in Table 2 and hardly interpretable in a specific direction.

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

Table 2 *Summary data*

	1985	1986	1987	1989	1992
Reading	–0.042	0.523		0.044	
English	0.647		–1.25		–0.538

6.2.3 Specific aspects of learning

The final category of studies (1, 4, 6, 7) looks at one specific aspect of learning (*writing, reading, spelling*) and in some cases gives a further breakdown as specified above. These studies are impressive in terms of methodology, taking great care to avoid confounding variables. This appears to be reflected in the fairly similar findings reported in these four studies, despite the fact that they varied significantly in size and scope and are unlikely to include the same primary studies (see Table 3).

Turning to the types of variables under investigation, it is not surprising that the large majority concern *writing* and *reading* because outcomes here are easy to measure numerically. However, it needs to be kept in mind that some of the studies (marked #) also contain other investigations which are not reported in our discussion. The smallest but most tightly controlled study in our sample (7) concentrated on *spelling*. We will now report findings separately for those variables.

6.2.3.1 Reading

The two studies dedicated to investigating *reading* (4, 6) report relative effect sizes of 0.13 and 0.25, i.e. while positive, not educationally significant (see Figure 1). These findings are supported by study 2 which reports an effect size of 0.26 for *reading*. Findings in study 8 are at variance with these, ranging from very low to very high (0.06, 0.43, 0.84). One explanation for this discrepancy may be related to the oddity of reporting median values (see discussion of study 8 above). This, however, does not account for the very high average value of 0.84 related to *writing-based reading programs*.[2] While the author does not comment on this specifically, it might be speculated that technologies especially dedicated to improve one specific aspect of learning, in this case writing programs designed to improve reading, are more useful. The same might apply for *reading management*, but in the absence of more detailed information of what that entailed, we are reluctant to draw firm conclusions from these results. It is important to keep in mind, though, that study 8 includes the largest number of subjects, exceeding that of the three previous studies. It is interesting to note that an earlier study by Kulik and Kulik (1991) confirms the trend reported here for effect sizes in relation to reading

2. Writing-based reading programs. Writing to Read (WTR) is a program that attempts to teach young children to read by stimulating them to write. The program is based on the premise that young children can learn to write in some fashion whatever they can say, and having learned to write, the children can then learn to read what they and others have written. The IBM Corporation became a major developer of the program during the 1980s, and in 1984 IBM released the first version of WTR. IBM released the current version of the program, called Writing to Read 2000, in 1991 (Kulik, 2003:v).

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

(overall average 0.25). We have not included this work because only seven studies out of a total of 254 were concerned with reading – the study mainly set out to investigate mathematics and science. It is worth noting, however, that they report an overall effect size for all the studies of 0.30, which reflects the most consistent findings here.

For a more comprehensive picture, we will include findings related to reading reported in Dunkel (1991). While it needs to be kept in mind that these refer to one study only, that of Pederson (1987), Dunkel recommended this particular approach as a design model. Results are therefore of interest to researchers wanting to compile a detailed synthesis on L2 reading, as recommended below.

> In an experiment on the cognitive processing of college students during computer-assisted L2 reading, Pederson examined three independent variables related to levels of cognitive processing during CALL reading practice in French: Verbal ability (low versus high), type of comprehension question (literal versus inferential), and passage availability during testing (reinspection of passage or lack thereof). Analysis of the protocols indicated that subjects recalled more information if they answered questions about a passage when it was unavailable for reinspection. Pederson inferred that the careful reading of the text in the passage-unavailability condition required readers to engage in self-interrogation, self-measurement of degree of comprehension prior to advancing the program, and auto-organization of the passage. Pederson's investigation demonstrated that use of a particular type of computer coding (e.g. passage availability) can affect achievement in a computer-assisted reading environment. Her study provided concrete, and potentially enlightening, data concerning the effective design of CALL reading materials (Dunkel, 1991:19).

6.2.3.2 Writing

Findings related to aspects of writing were more consistent on the whole, although we must point out that the same problems are present in study 8 with reporting median values. It is interesting that both authors have taken the trouble to report separate effect sizes for subcategories common in the primary studies included. Study 1 reports the effect of word processing on L1 writing. Effect sizes were calculated for five outcome measures: *quality of writing* (0.27), *number of words produced* (0.52), adhering to writing conventions – e.g. correct punctuation, verb/subject agreement – (0.16), *frequency of revision* (0.18) and *attitude towards writing* (0.12). In the last three measures the effect size for each study ranged widely from negative to positive and the author points out that, in each case, the median was very close to zero. The author also signals that a very large effect size in one study inflated the result for number of words produced, and that the median (0.36), an indication of central tendency, may be more meaningful in this case.

Study 8 reports findings for *writing using word-processing* which corresponds to the overall category in study 1; and *writing using computer enrichment*.[3] No effect sizes are

3. Computer enrichment. Schools provide students with enriched computer access in a variety of ways. They may let students work on a variety of relatively unstructured exercises in the classroom, including games, simulations, and tutorial programs. Or they may extend student access to Internet resources, or they may provide students with anytime-anywhere laptop access to computers. A common goal in enrichment programs is to help students develop as writers by giving them more opportunities for authentic writing and research. (Kulik, 2003:vii)

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

Table 3. *General overview of reviews*

Id	Year Published	Period encompassed	Ed Setting	Nº Studies	Nº Students	Type of Review*	Variables	Finding
9	1997	1987-1994	Tertiary	35		Interpretive summary	L2 Reading comprehension	Can substantially improve comprehension
							L2 Communication	Creates opportunities for conversation with native speakers
							L2 Feedback	Type of feedback important – different strategies can result in different outcomes
							L2 Student attitude	Attitude not consistently linked to achievement
							L2 Achievement	3 of 4 studies found significant gains.
10	1998	20 years	Primary, Secondary & Tertiary	NA	NA	Literature review	L1 & L2 Reading	Has the potential to actively engage students. Holds great promise for becoming a powerful tool.
11	2001	1982-1997	Adult education	23	>1679*	Analytical review	L1 Literacy	Some evidence supporting use of computers for adult literacy. A higher rate of success for students with reading level of Grades 4-6.
12	2001	1990-2000	Kindergarten & Primary	5	373	Qualitative research synthesis	L1 word learning	3 statistically significant. 1 very positive. 1 positive after data manipulation. 1 negative.

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

Table 3. *General overview of reviews (con't)*

Id	Year Published	Period encompassed	Ed Setting	Nº Studies	Nº Students	Type of Review*	Variables	Finding
13	2002	1990-2000	Primary, Secondary & Tertiary	246, 70 research-based	NA	Literature review	L2 Writing using CMC	Increased participation; Positive attitudes; greater student empowerment; decreased teacher control; wider variety of discourse functions; mixed results regarding syntactic complexity.
							L2 Reading	Computer supported glossing formats helpful in developing proficiency; visual annotations for words greatly enhances vocabulary learning.
							L2 Speaking	Speech recognition software shows promise but currently not satisfactory. Computer books provide the content on which discussion could centre.
							L2 Listening	Multimedia offers different modalities and therefore may cater more effectively to different learning styles.

*as defined by the authors

reported for *writing using computer prompts*.[4] All effect sizes show a positive trend hovering around the 0.3 mark, i.e. approaching the level regarded as educationally beneficial by most researchers in this field (see Figure 1). Seeing that study 1 can be regarded as exceptionally tightly controlled in terms of design and that it included large numbers of primary studies (33) and subjects (>1874), the reported effect size (0.27) for *quality of writing* can be viewed with some confidence. The fact that study 8 reflects these findings might lend further support to these results, but in the absence of the range of effect sizes, we cannot be certain. Study 1 also strongly suggests that students using word-processing tools produce more words, though of course this in itself does not suggest better writing quality.

6.2.3.3 Spelling

The study investigating this very specific variable (**7**) is another excellent design model included in the discussion below. While it is the smallest study in terms of the number of primary investigations and subjects involved (7, 240), the findings carry high confidence levels. Interestingly, the reported overall effect size (0.37) is quite similar to those found for reading and writing and even slightly higher overall. However, the rigorous statistical analyses applied in this study revealed that this result fell just below significance level (p=0.06) and the authors are therefore very cautious in their interpretation:

> This review suggests that the teaching of spelling by using computer software may be as effective as conventional teaching of spelling, although the possibility of computer-taught spelling being inferior or superior cannot be confidently excluded due to the relatively small sample sizes of the identified studies. Ideally, large pragmatic randomised controlled trials need to be undertaken. (Torgerson & Elbourne, 2002:129)

In the context of our total pool of findings and considering the fact that results would have been significant at p=0.05, this interpretation is perhaps a little too cautious. It lends support to the observations made in Felix (2005) which suggested an inverted correlation between soundness of design and confidence in reporting results.

In summary, what can be concluded with some confidence from the meta-research reported here is that the use of educational technologies appears to have a small positive effect on (in order of magnitude) L1 spelling, L1 writing quality and fluency and L1 reading.

6.3 Reviews

Because of the very wide-ranging scope of the studies in this sample, it is very difficult to produce succinct summary findings. The information in Table 3 gives some impression of the sorts of variables under discussion and synthesized findings as reported by the authors.

---

4. Computer writing prompts. Researchers have modified standard word processing programs to provide writing prompts while students are composing with the programs. The prompts may focus on writing mechanics, or they may provide guidance in development of new ideas. (Kulik, 2003:vii)

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

Three studies deal specifically with CALL while the other two investigate *L1 reading, literacy* and *word-learning*. Approaches vary considerably from straight literature reviews reporting results discursively with minimal interpretation (10) to qualitative research syntheses including similar procedures to meta-analyses but without calculating effect sizes (12). It is interesting to note that the principal authors of the latter also carried out study 6 discussed above. While they had set out to calculate effect sizes, they decided that the amount of data contained in the five primary studies did not warrant such analyses. This lends some support to our critique of study 5 above and Dunkel's (1991) rejection of Roblyer *et al.*'s (1988) findings.

Studies investigating results in L1 environments concentrate on one specific variable, while two of the L2 studies report a large variety of aspects, either broken down into common variables identified in the primary studies (9), or reported under the headings of the traditional four language skills (13). While this adds great detail to the general discussion, it gives the impression of a large unruly pool of data, in which trends appear difficult to ascertain. Seeing that findings pertain to even more variables than those reported in Dunkel (1991), this body of work on its own cannot be expected to clarify Dunkel's equivocal findings. What these studies add to our discussion in qualitative terms, however, is very useful in three distinct ways. (1) They give limited support to the findings reported above related to reading and writing; (2) they flesh out what has been said in quantitative terms by adding more detailed observations; and (3) they give a good impression of significant gaps in the research on effectiveness. The vast data collected in these studies, especially in 9 and 13 represents an excellent starting point for the sorts of investigations recommended at the end of our paper. For instance, there are some findings related to the very important variables *feedback* and *speech recognition* that warrant further investigation.

In summary, tentative conclusions that can be reached here are that findings on the whole are positive, although it needs to be kept in mind that until recently negative results were rarely reported in the literature. More specifically, it has been found that technologies have the potential to engage students and create opportunities, adding value to face-to-face instruction; that dedicated programs (such as glossing and visual annotations for word-learning) are useful; that the multimodal nature of current technologies appeals to different learning styles, that the use of technologies can have a positive effect on student attitudes and participation (although not reflected in higher achievement), and that L1 literacy benefits from the use of ICTs. In terms of achievement, these reviews reported significant positive results in four out of seven studies. It is interesting to note that the authors here also express concerns related to the quality of research designs and validity of findings.

7 Design models

Four studies in our sample can be singled out as good-practice models for meta-analyses (1, 4, 6, 7). All of these concentrate their efforts around one particular aspect of learning which makes interpretation of results much more meaningful than the overall effect sizes reported in *mega* studies (i.e. exceptionally large meta-analyses) dealing with a large variety of subjects/variables and settings. While the latter doubtlessly contribute useful data and background to the debate about effectiveness in general, they are less

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

valuable in informing good teaching practice, choice of technologies, or even further research. Since one of the biggest concerns about this type of research, mentioned by most authors in our sample, relates to the differential measures of control applied in primary studies, confidence levels about overall findings cannot be very high. What can be hoped, though, is that when all research – *individual, meta* and *mega* - is looked at together, trends might emerge that can be interpreted with some degree of confidence. As we have seen above, exceptional findings can sometimes be traced back to poor research methods and/or reporting of results. In the following we will briefly discuss the sorts of steps that the authors of two of the models (1, 7) have taken in order to ensure maximum confidence levels and meaningful interpretation of results.

In a meta-analysis it is important to document fully the method for searching the literature and the rules for inclusion of studies (Schafer, 1999). Both studies 1 and 7 give a well-documented search strategy and very comprehensive and detailed inclusion criteria. While both studies concentrated on one major variable, Study 1 calculated effect sizes for five subcategories (see section 6.2.3.2 on *writing* above). Study 1 also coded each primary study for 21 characteristics under four categories: instructional treatment, research methodology, study setting and publication features. These characteristics act as moderators for the effect size and enhance generalizability (Schafer, 1999). For example, while the study reported an overall effect size of 0.27 for writing quality, the authors then went on to show how the effect size changed according to certain study characteristics – where subject assignment was non-random the effect size was 0.39, where it was random it was 0.31, and where student groupings were voluntary it was 0.03. This type of information is somewhat more informative than simply giving an overall positive effect size.

Both studies include tables listing each primary investigation, the sample size, setting duration, effect size and a number of study characteristics, as recommended by Slavin (1986). Such a table provides a useful summary and makes it easier to check the findings against the original research.

Study **7** calculates both a confidence interval and the significance for each effect size. Study **1** calculates statistical significance and reports the *Standard Error* for each effect size. When reporting the findings, both the mean and the median effect size, together with an indication of range are given. Attention is drawn to any outliers.

8 Concluding remarks and suggestions for future research

Revisiting the question posed in the title of our paper, we have to conclude that 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. However, the inclusion of narrative reviews and studies carried out in L1 learning environments helps to clarify the picture somewhat. What emerged very clearly is that effectiveness research in L1 and L2 learning since the 1980s has largely focused on variables related to reading and writing and to a lesser extent on word learning. This, of course, is not surprising since it is relatively easy to measure outcomes related to this type of learning. Small positive gains are reported consistently; most of these, however, fall below significance level. The most conclusive finding in this sample relates to *L1 writing using word-processing tools,* where effect sizes approach a level that can be

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

interpreted as educationally beneficial (a move from the 50th to the 62nd percentile for the median student). Results for *writing fluency* are even higher. Findings related to reading are more mixed, with the most rigorous studies reporting similar effects to the above but higher and lower values are being reported in another study. This last study also suggests that dedicated writing tools may improve *L1 reading*. While findings of this study need to be read with caution because of procedural concerns, they are derived from exceptionally large data pools. Observations reported in the narrative reviews lend some support to our conclusions derived from the meta-analyses and provide a little more detail on the types of areas investigated related to writing and reading.

Overall, there is something to be said for looking at large amounts of data, especially in an area that is dominated by small-scale studies with small numbers of subjects. Even if some research methods in this sample are lacking in rigor, the authors supply data that might be usefully interpreted in the light of work carried out elsewhere and recommended below. They also generally provide excellent literature reviews. There appears little value, however, in reporting overall effect sizes when the pool of investigations includes too much variety in terms of variables, settings, methods and technologies used. Ideally, the role of meta-analyses would be to provide us with a rigorous synthesis of data related to a particular question of interest. After all, that is the specifically stated purpose of such research as compared to a narrative review. In our sample, only four studies did this effectively. While one of these, investigating the single variable *spelling*, contained the least number of studies (7) and subjects (240), it was carried out in as rigorous a fashion as can be hoped in this kind of research. We can therefore be reasonably confident that *L1 spelling* potentially benefits from the use of ICT.

It is, however, advisable to keep in mind the authors' caution in terms of interpreting results in most studies in our sample. Again, as previously reported in Felix (2005), we observed a correlation between excellence of research design and caution exercised in reporting results and vice versa. In any case it is well to remember that even when results are derived through sound procedures and analyses, there may still be other interesting explanations for the positive results. Bangert-Drowns (1993:74) points out a pertinent observation by Russell (1991) who investigated effectiveness of word-processing:

> '…in cases where word processing seemed beneficial, the benefits may not be due at all to the word-processing itself but to the kinds of social interactions that computer laboratory environments permit.'

In educational terms, of course, it hardly matters what exactly produced the improvement and, if we follow the cause and effect argument to its logical conclusion, we will find that the technology itself is the least likely contributor (see also Clark, 1983; 1985; 1994). Teachers, programmers, methods, settings, social aspects and learning processes in ICT environments are likely to contribute their inestimable share. Therefore, building up an even larger body of data, including these sorts of aspects, by means of complementary research designs generating sound qualitative and quantitative data may produce trends - the beginnings of which we have observed here – that may be interpreted with the type of confidence levels sorely needed in this controversial area.

In conclusion, our systematic look at the research – starting with Dunkel (1991), examining recent work (Felix 1995) and finally filling in the gaps by investigating the findings

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

of meta-analyses and narrative reviews – has clarified the picture about effectiveness in several ways. It has (1) made us understand better the difficulties associated with carrying out research in this important field; (2) furnished us with situated good-practice research models; (3) allowed us to come to a small number of conclusions with some confidence; and (4) enabled us to formulate suggestions for research which will build on this knowledge. Our recommendations for future work in this area are as follows:

1. Rigorous meta-analyses of the type discussed in the models above, would contribute useful quantitative information, especially in the light of new variables emerging in recent social constructivist learning contexts, such as the role of *collaboration, meta-cognitive skills and knowledge* or *online presence* and *identity*.
2. Qualitative and discursive syntheses of a body of research investigating similar variables related to one larger issue such as *reading*, for instance, would provide comprehensive and detailed data not hitherto available.
3. Further high-quality, single experimental and non-experimental studies of areas relatively unexplored, such as *speaking online*, would add important new data.
4. Replication studies of excellent previous work would strengthen existing data, especially if larger subject pools could be found.

Carrying out this sort of work by way of an agreed research agenda and disseminating results as soon as they came to light, would advance this controversial field tremendously and might help to avoid the notorious re-invention of the wheel. Our Research Centre for New Media in Language Learning has made a modest start in this direction, and we would welcome co-operation from interested colleagues world-wide.

Acknowledgements

My grateful thanks to Nina Dow for her contribution to the data gathering and compilations and to Lesley Shield for her thoughtful feedback on our first draft.

References

Bangert-Drowns, R. L. (1993) The Word Processor as an Instructional Tool: A Meta-Analysis of Word Processing in Writing Instruction. *Review of Educational Research*, **63**(1): 69–93.

Chapelle, C. C. and Jamieson, J. J. (1991) Internal and External Validity Issues in Research on CALL Effectiveness. In: Dunkel, P. (ed.), *Computer-Assisted Language Learning and Testing*. New York: Newbury House, 37–59.

Christmann, E. P. and Badgett, J. L. (1997) Progressive Comparison of the Effects of Computer-Assisted Instruction on the Academic Achievement of Secondary Students. *Journal of Research on Computing in Education*, **29**(4): 325–337.

Christmann, E. P., Badgett, J. L. and Lucking, R. (1997) Microcomputer-Based Computer-Assisted Instruction Within Differing Subject Areas: A Statistical Deduction. *Journal of Educational Computing Research*, **16**(3): 281–296.

Clark, R. E. (1983) Reconsidering Research on Learning from Media. *Review of Educational Research*, **53**(4): 445–459.

Clark, R. E. (1985) Evidence for Confounding in Computer-Based Instruction Studies: Analyzing the Meta-Analysis. *Educational Technology Research and Development*, **33**(4): 235–262.

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

Clark, R. E. (1994) Media Will Never Influence Learning. *Educational Technology, Research and Development*, **42**(2): 21–29.

Cooper, H. M. and Hedges, L. V. (1994) Potentials and limitations of research synthesis. In: Cooper, H. M. and Hedges, L. V. (eds.), *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, 521–529.

Dunkel, P. (1991) The effectiveness of research on computer-assisted instruction and computer-assisted language learning. In: Dunkel, P. (ed.), *Computer-Assisted Language Learning and Testing*. New York: Newbury House, 5–36.

Felix, U. (2005) Analyzing recent CALL effectiveness research: Towards a common agenda. *Computer Assisted Language Learning,* **18** (2&3): 1–31.

Glass, G. V., McGaw, B. and Smith, M. L. (1981) *Meta-analysis in Social Research*. Beverly Hills: Sage Publications.

Hope, G., Taylor, H. and Pusack, J. (1984) *Using Computers in Teaching Foreign Languages*. Orlando, FL: Harcourt Brace Jovanovich.

Kulik, J. A. (2003) *Effects of Using Instructional Technology in Elementary and Secondary Schools: What Controlled Evaluation Studies Say* (No. P10446.001). Arlington, VA.: SRI International.

Kulik, J. A., Kulik, C. C. and Cohen, P. A. (1980) Effectiveness of computer-based college teaching: A meta-analysis of findings. *Review of Education Research*, **50**(4): 525–544.

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.

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

Miech, E. J., Nave, B. and Mosteller, F. (1997) On CALL: A Review of Computer-Assisted Language Learning in U.S. Colleges and Universities. *Educational Media and Technology Yearbook*, **22**: 61–84.

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 Co, 99–131.

Research Centre for New Media in Language Learning
http://www.arts.monash.edu.au/lcl/newmedia_in_langlearn/index.htm

Roblyer, M. D., Castine, W. H. and King, F. J. (eds.) (1988) *Assessing the Impact of Computer-based Instruction: A review of recent research.* New York: Haworth Press.

Russell, R. G. (1991) A meta-analysis of word processing and attitudes and the impact on the quality of writing. *Annual Meeting of the American Educational Research Association*, Chicago, IL.

Schafer, W. D. (1999) Methods, plainly speaking: An overview of meta-analysis. *Measurement and Evaluation in Counseling and Development*, **32**(1): 43–61.

Singhal, M. (1998) *Using Computers as Reading Instructional Tools: Applications and Implications*. http://www.eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/25/69/0f.pdf

Slavin, R. E. (1986) Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. *Educational Researcher*, **15**(9): 5–11.

Suri, H. (2000) A Critique of Contemporary Methods of Research Synthesis. *Post-Script*, **1**(1): 49–55.

Zhao, Y. (2001) Recent developments in Technology and language learning: A literature review and meta-analysis. *CALICO Journal*, **21**(1): 7–27.

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press

Further reading: Meta-analyses and Reviews by Id

Bangert-Drowns, R. L. (1993) The Word Processor as an Instructional Tool: A Meta-Analysis of Word Processing in Writing Instruction. *Review of Educational Research,* **63**(1): 69–93.

Christmann, E. P., Badgett, J. L. and Lucking, R. (1997) Microcomputer-Based Computer-Assisted Instruction Within Differing Subject Areas: A Statistical Deduction. *Journal of Educational Computing Research*, **16**(3): 281–296.

Christmann, E. P. and Badgett, J. L. (1997) Progressive Comparison of the Effects of Computer-Assisted Instruction on the Academic Achievement of Secondary Students. *Journal of Research on Computing in Education*, **29**(4): 325–337.

Soe, K., Koki, S. and Chang, J. M. (2000) Effect of Computer-Assisted Instruction (CAI) on Reading Achievement: A Meta-Analysis. Retrieved 18/2/2004, from http://www.prel.org/products/Products/effect-cai.htm

Blok, H., Oostdam, R., Otter, M. E. and Overmaat, M. (2002) Computer-assisted instruction in support of beginning reading instruction: A review. *Review of Educational Research*, **72**(1): 101–130.

Singhal, M. (1998) *Using Computers as Reading Instructional Tools: Applications and Implications*. http://www.eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/25/69/0f.pdf

Torgerson, C. J. and Elbourne, D. (2002) A Systematic Review and Meta-Analysis of the Effectiveness of Information and Communication Technology (ICT) on the Teaching of Spelling. *Journal of Research in Reading*, **25**(2): 129–143.

Kulik, J. A. (2003) Effects of Using Instructional Technology in Elementary and Secondary Schools: What Controlled Evaluation Studies Say (No. P10446.001). Arlington, Virginia: SRI International.

Miech, E. J., Nave, B. and Mosteller, F. (1997) On CALL: A Review of Computer-Assisted Language Learning. In: U.S. Colleges and Universities. *Educational Media and Technology Yearbook*, **22**: 61–84.

Berger, J. (2001) Effectiveness of Computers in ALBE Classrooms: An Analytical Review of the Literature. *Adult Basic Education*, **11**(3): 162–183.

Blok, H., van Daalen-Kapteijns, M. M., Otter, M. E. and Overmaat, M. (2001) Using Computers to Learn Words in the Elementary Grades: An Evaluation Framework and a Review of Effect Studies. *Computer Assisted Language Learning*, **14**(2): 99–128

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

https://doi.org/10.1017/S0958344005000923 Published online by Cambridge University Press