ReCALL **20**(2): 141-161. 2008 © European Association for Computer Assisted Language Learning
DOI: 10.1017/S0958344008000323 Printed in the United Kingdom

The unreasonable effectivness of CALL: What have we learned in two decades of research?

USCHI FELIX
School of Languages, Cultures and Linguistics,
Monash University, Melbourne, Australia
(*email: uschi.felix@arts.monash.edu.au*)

Abstract

This paper presents a comprehensive picture of what has been investigated in terms of CALL effectiveness over the period 1981-2005 throwing light on why this question is still such a difficult one to answer unequivocally. The author looks at both strengths and weaknesses in this body of work, highlighting pitfalls and paradoxes in research procedures and providing valid design models. This includes the contribution of dedicated meta-analyses to this controversial field and a discussion of the benefits and limitations associated with this type of research. Substantial data, drawn from three extensive studies (Felix, 2005a, b; Felix, 2006a), allows the author to present for the first time synthesized findings relating to the impact of technologies on language learning. The paper concludes with strategies for future work in the context of a proposed research agenda.

Keywords: CALL effectiveness, meta-analyses, CALL research agenda

Preamble

In order to dispel some readers' dislike for numbers and statistics, we draw attention to this very encouraging quote from the great Bertrand Russell (in Dennon & Egner, 1992:253):

> Mathematics, rightly viewed, possesses not only truth, but supreme beauty, a beauty cold and austere, like that of sculpture, without appeal to any part of our weaker nature, without the gorgeous trappings of painting or music, yet sublimely pure, and capable of a stern perfection such as only the greatest art can show. The true spirit of delight, the exaltation, the sense of being more than Man, which is the touchstone of the highest excellence, is to be found in mathematics as surely as in poetry.

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

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 language 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.

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 closely to the project's context, methods and analyses. Curiously, all of this has already been done in one way or another over the past two decades, yet general conclusions and claims remain largely equivocal.

In order to provide a clearer picture, we spent the last four years compiling and analyzing a vast amount of data related to CALL effectiveness published in English. Because of the overwhelming volume of papers located in our searches, including articles written in other languages was at this stage beyond our capacity. We began by examining the entire meta-research conducted in the field (Felix, 2005a). While this provided us with a little more certainty regarding statistical outcomes that might be generalised, it did not dispel frustrating worries about the validity of some research designs. Clearly, more close attention needed to be given to this.

Our next study (Felix, 2005b), therefore, was dedicated to examining the sorts of designs researchers had used to ascertain effectiveness. The study highlighted strengths and weaknesses in this body of work and singled out models of good design practice. In this context, the usefulness of meta-analyses to determine clear cause and effect results, relying solely on effect sizes, hence measuring *outcomes* in numerical terms, appeared questionable. It became clear that a series of systematic qualitative syntheses of findings related to one particular variable such as *learning strategy* or *writing quality* might produce more valuable insights into the potential impact of technologies on *learning processes* as well as *outcomes*.

The next step was to test this assertion, and our final project (Felix, 2006a) was designed to look closely at all studies which had dealt with L2 writing. It was hoped that this would add important qualitative information, complementing the quantitative findings of the meta-research, especially since the latter had largely concentrated on L1. We expect that drawing conclusions in the light of this extensive body of work 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

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

area of interest and thus strengthen their work. To save space we have not included the extensive Tables, References and Appendices generated for the three studies. The reader is referred to our Research Centre Website – http://www.arts.monash.edu.au/lcl/newmedia-in-langlearn/ – should this very detailed information be required. In the following we will:

- define what is meant by *effectiveness, meta-analysis* and *effect size*
- outline the scope of the three previous studies on which our discussion is based
- identify strengths and weaknesses of the *effectiveness* research
- single out successful research design models
- highlight valid meta-analysis procedures
- present a synthesis of major findings
- draw conclusions and suggest directions for future research.

2 Definitions

2.1 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 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.

2.2 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)

2.3 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 means of a dependent variable at the end of a treatment phase. See Figure 1 for an accessible, comprehensive description by Miech, Nave and Mosteller (1997).

https://doi.org/10.1017/S0958344008000323 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

(Miech et al., 1997:76)

Fig. 1. Effect sizes

3 Scope of the previous studies on which our findings are based

3.1 Study 1

What do meta-analyses tell us about CALL effectiveness? (Felix, 2005a). The literature search for this study was restricted to papers published after 1991. The reason for the cut-off was that these studies included investigations going back as far as 1980 and that Dunkel (1991) gives a comprehensive overview of reported effectiveness in this earlier work. Our searches resulted in more than 150 papers. Most of these were discarded because they were not concerned with *effectiveness* as defined or were not research reviews. This left 13 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 narrative 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 were relevant to CALL. Overall several hundred research studies were contained in these 13 meta-analyses and reviews with the number of subjects exceeding 20,000.

3.2 Study 2

Analyzing recent CALL effectiveness research: Towards a common agenda (Felix, 2005b). Since the purpose of this study was to take a detailed look at strengths and weaknesses of research designs, our search had to be limited in scope. It was decided to examine work published during the years 2000-2005. To be included, studies needed to report the results of research about the effectiveness of the use of ICT on language

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

learning processes or outcomes in the widest sense. Because our goal was to give a comprehensive overview of types of research conducted, we did not cull these studies to include only those with sound research designs. These searches resulted in more than 150 studies. Many of these were discarded because they were concerned with learning in areas other than CALL or did not investigate an aspect of effectiveness of ICT on language learning. The final number of studies included here was 52.

3.3 Study 3

Analyzing the impact of educational technologies on L2 writing (Felix, 2006a). This study examined research (1991-2005) where students were engaged in a writing task and where either the writing was subjected to some form of evaluation, or some aspect of the writing process was examined. In some studies the writing may have been evaluated for one or more aspects such as grammatical accuracy, appropriate use of vocabulary, language functions and register, or correct use of structure/paragraphing. In others, the research was more concerned with processes such as revision strategies, participant roles, or the degree of acceptance of different modes of feedback.

Our search returned 192 papers. Many of these were rejected because (1) they were not research studies; (2) the writing output was not evaluated; (3) writing was only a very minor part of the study; or (4) computers were used to do the research, not instrumental in the teaching or learning. This left a total of 62 studies which were analysed.

3.4 Categorization in all three studies

Our aim was to provide the widest possible breadth of information on the research included. Because of the great variation in the many characteristics of each study, this proved tremendously difficult, especially where variables under investigation were concerned. In order to present so much information in an intelligible format we sometimes had to group several variables under an umbrella category. For example, many studies looked at different aspects of *reading* (close reading, critical reading, reading comprehension etc). These were grouped together as *reading skills*. The major categories chosen were:

- Number of participants
- Research design used
- Technology used
- Educational setting
- Language taught
- Subject/skill taught
- Variable under investigation
- Reported findings.

We would have liked to include *instructional method* and context (i.e. how well supported the use of ICT was in the particular setting), but information on these very important elements remains very scarce indeed. (see also Hubbard, 2004; Jamieson & Chapelle, 2004; Levy, 2004).

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

4 Overview and strength of the effectiveness research

Taking into account earlier studies (see Basena & Jamieson, 1996; Dunkel, 1991; Soe *et al*, 2000) we are beginning to see a substantial body of qualitative and quantitative data related to CALL effectiveness. While concerns expressed by Phipps and Merisotis (1999) regarding the lack of random assignment of subjects and control for extraneous variables are still present, they are no longer sustainable to the same extent. We are finding an increasing number of sound quasi-experimental and experimental studies, a substantial number of which include randomly selected control groups.

In our sample the distribution between experimental (including quasi-experimental) and non-experimental (including pre-experimental) studies is almost equal (definitions can be found in the Appendix). This is indeed a strength of the research since both categories include a large number of excellent studies and their combined results advance the field more substantially than a continuous series of highly controlled studies. This is especially the case since the latter are predominantly very short-term (as brief as a 40 minute treatment) and most often investigate a single variable.

As predicted by Dunkel (1991), there is a noticeable move towards the inclusion of *learning processes* (see also Collentine, 2000). This is not surprising since teaching methods and technologies have changed in a way that no longer allows for easy numeric measurement of outcomes. Overall there is a distinctly higher percentage of investigations concerned with quality and process measures than with quantity. This trend is also reported by Hoven (2004). If we look at all the studies that investigate writing, for instance, more than half deal with writing quality and are spread fairly evenly across the spectrum of research designs.

Although investigations in tertiary settings still dominate the field – also reported by Liu *et al* (2002) who reviewed CALL studies pre 2000 – we are seeing an increased number of studies emerging in the school environment (11.5% in our sample). This, too, lends strength to the body of research since it provides important opportunities for investigating potentially differential impacts of ICT on learning in elementary, secondary and tertiary settings.

We are also beginning to see languages emerging other than those that have always been prominent in the field in CALL, i.e. Japanese, Indonesian, Chinese and Russian. Interestingly the distribution appears unchanged from when we looked at Web-based resources (Felix, 2001) with ESL/EFL, Spanish and French far outweighing any other language. However, our data show a useful spread of languages included in investigations of many variables.

5 Common problems in the effectiveness research

5.1 Misleading titles

Many studies claim to be investigating effectiveness of a form of ICT but are really looking at its viability as a teaching and learning environment on the basis of surveys of student perceptions. Titles can often be very misleading when they include the keyword effectiveness. For example, the otherwise interesting study by Dolle and Enjelvin (2004) is entitled Investigating *"VLE-ffectiveness" in languages*, when a more representative title would have been *Investigating student perceptions of a VLE as a viable learning*

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

environment. Since no measurements or comparisons of learning outcomes were carried out, the study cannot produce information on "key factors in enhancing learning effectiveness in a VLE" (*op. cit.*, 2004:486), as the authors claim. Instead it gives valuable insights into the students' perceptions of the teaching and learning environment effected and affected by the technologies. It is one of the few studies that describe in detail the instructional methods and the technologies used. In studies like these it is important to remember (and the authors themselves list a number of limitations) that students' statements such as "formally assessed Web-based tasks have improved my overall coursework grade" (*ibid*, 2004:485) cannot be meaningful in the absence of any form of controlled comparison. The fact that the statistical analyses established a significant relationship between this statement and the one that said "assessed Web-based tasks encouraged me to engage more in the learning process" (*ibid*, 2004:485), does not establish enhanced learning effectiveness but that students think they are doing better. Although this information is useful and interesting, there is no way of knowing whether a different kind of intervention would not have produced the same result, even without possible Hawthorne and Pygmalion effects (see Appendix for definition) being at play.

5.2 Poor description of the research design

A thorough description of the procedures, including information on the subjects, materials, technologies, treatments, tests, statistical analyses and anything else pertinent to the particular investigation is absolutely essential in rigorous research (Nutta *et al*., 2002 is an excellent example here). However, this is far from standard practice, and it is sometimes surprising how studies in which procedures are not fully explained get published. For example, a great limitation of the otherwise thorough study by Chikamatsu (2003) is the lack of information about the subjects. In a design where the effects of computer use on Japanese writing is investigated, information on students' previous experience of computer-assisted writing, as well as their IT literacy, is crucial. The study also gives no background as to current or previous teaching methods and uses of CALL in the context of comparing handwritten and computer written tests when some or all of these constitute potential threats to internal validity. Another study (Myers, 2000) provides so little information on the research design of a study reporting on the effects of voice recognition software on perception and speech production in second language use, that it was impossible to fit it into our classifications grid.

5.3 Failure to investigate previous research

A surprisingly large number of studies do not begin with a thorough investigation of what has already been done in the area to be studied. Notable exceptions are the excellent review on writing and CALL in Chikamatsu (2003) and reading and CALL in Soe *et al* (2000). This often leads to duplications of efforts when building on and incorporating existing findings would advance the field more substantially. It is most surprising to find a statement such as the following when investigating student perceptions of CALL represents the largest area of activity to date, and especially at the time of this report.

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

> 'In spite of the widespread use of computer-assisted language learning (CALL) and its perceived facilitative role in second language (L2) learning, there is little data on how learners feel, experience, or think about CALL in the L2 learning context' (Suh, 2002:669).

5.4 Poor choice of variables to be investigated

There are a number of studies comparing excellent CALL activities with poor non-CALL exercises or inferior CALL. This is especially disappointing in studies where much effort has been invested in designing well-controlled procedures. An interesting case in point here is the project by Yeh and Lehman (2001) which we had singled out as one of the best experimental designs in the current body of work. Among others the study looked at the important variable of *learner control* in CALL, but what was compared was a system that allowed students a great deal of freedom in viewing and repeating video and text segments in any order and checking a glossary with a system in which pace and sequence were completely fixed and repetition was not possible. The study simply confirmed what we already know through a relatively large body of literature, when the rigorous design would have allowed for more interesting probing of learner control.

5.5 Overambitious reporting of results

Many studies still lack a discussion of limitations which, in an environment where perfectly controlled designs are near impossible, is most surprising. There is a definite trend towards excellent designs being accompanied by detailed discussions of possible confounding variables and cautious reporting of findings (even overly cautious as in 6.2 below) and for poor designs not recognizing threats to internal or external validity and reporting results in very certain terms, such as "this study proves" or "as has been demonstrated / shown", and so on. Since numerous books and articles have been written on the subject, we will not include here yet another outline of potential threats to the validity of findings but refer the reader to the detailed discussion in Chapelle and Jamieson (1991). A simple check of what other than the particular treatment might have an influence on results would take researchers a long way in setting up studies and writing up results realistically.

6 Design models

6.1 General comments

Because there is such a large scope for research in this area, there cannot be a single best design model. What is imperative, though, is that researchers match the design to the research questions, the context in which the study takes place, the time-frame available, the variables under investigation, their capacity for statistical analyses and their ability to control for confounding elements. A short-term fully controlled experimental design, for instance, would be suitable to measure individual well defined outcome effects (see 6.2 below), while a longer-term non-experimental study using qualitative measures such as observational procedures and think-aloud protocols would yield important data related to effects on learning processes. A combination of various data collection

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

methods within one single study will help in strengthening confidence levels about results (see 6.3 below).

We have singled out below examples of good design practice in four distinct and important areas including both outcomes and processes: (1) a study of differential effects of multimedia elements on vocabulary learning; (2) an investigation of the effectiveness of CALL on language proficiency; (3) a project examining whether threaded discussions can be effective in realizing constructivist principles; and (4) a case study examining the role of visually rich technology in facilitating children's writing.

6.2 Within-subject design

Al-Seghayer (2001) used a within-subject design (n=30) to test the effect of multimedia annotation on vocabulary acquisition under three conditions: printed text definitions; printed text definitions linked to still pictures; and printed text definitions linked to video clips. Data was gathered in a variety of complementary modes. In qualitative terms, students were asked to complete a questionnaire and take part in a face-to-face interview, giving their own view of which condition might be most conducive to vocabulary learning or conveying of meaning. In quantitative terms, a recognition and a production test were administered and data processed by analysis of variance procedures. Results of both qualitative and quantitative data led to the conclusion that video clips are more effective than still pictures in teaching unknown vocabulary.

This study represents a sound approach to the problem. It contains an extensive and useful literature review of related studies and sets out to fill a gap in the research. While the review gives strong evidence for the effectiveness of both still pictures and video clips in a variety of language learning activities, the new study represents the first attempt to compare the two modes experimentally.

A great deal of care was taken in controlling for confounding variables and in describing procedures in every detail. Each subject served as their own control by taking part in each of the conditions. The annotated items of vocabulary used in each condition were controlled for frequency, grammatical category, morphological category, visual complexity and for whether they represented abstract or concrete concepts. An unknown reading passage was chosen and adapted for intermediate ESL learners applying the criteria text length, syntactic complexity and content and pre-tested with subjects not taking part in the study.

The author outlines the following limitations: (1) The small sample sheds doubt on the validity of the observed significance; (2) Assessment of the learning outcome was measured only with multiple choice and production tests; (3) The study did not analyse individual performance data such as study path or reaction time; and (4) Only short-term retention was studied (Al-Seghayer, 2001:227). It would therefore be interesting to replicate this study with many more subjects, and specifically address these limitations. Another useful measure to add in a larger study would be to carry out a correlation analysis between the quantitative findings and the students' perceptions of the most effective mode. Could it be that they performed better in this mode because they believed they would?

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

6.3 Experimental design

The study by Nutta *et al* (2002) is an excellent example of a well-designed experimental investigation. It is one of the very few that was carried out over an extensive period of time (three hours by 25 weeks for the experimental treatment, 13 months in all including an equal instruction period and collection of qualitative data) and also one of a limited number conducted in an elementary school setting (n=28). The study set out to test the effect of multimedia materials on proficiency in Spanish. Students were assigned at random to a treatment and a control group with the same instructor carrying out the teaching in both. A great deal of care was taken in providing students with near identical activities.

The extensive quantitative measures included pre- and post-tests on proficiency, scored by an independent native speaker, and a criterion-referenced post-test on achievement, including oral and written items, developed by another independent researcher. The achievement test was checked for internal consistency of the written and oral sections combined and also re-administered in delayed mode. Qualitative measures were equally extensive, albeit limited to a smaller number of subjects for some procedures, and consisted of whole-group observations, interviews with students, teachers, administrators and parents and think-aloud protocols. Because of the nature of the qualitative investigations this study gives more detailed information on instructional methods and context than is usually found in experimental studies, especially those that are limited to a one-off treatment. The interviews with parents, for instance, supported the often voiced ethical dilemma in this sort of design in which students might be deprived of a potentially better learning environment. One parent noted "My child says that she does not get to use the computer" (Nutta *et al*, 2002:304).

No statistically significant difference was found at post-test between the experimental group and the control group (who had used printed and audio materials). A significant difference, however, was reported on the delayed post-test in favour of the treatment group. The qualitative investigations detected differences in language behavior, with the students who used multimedia spending more time to stop, check, and revise their language production, leading to greater precision in pronunciation and the use of larger chunks of language when repeating phrases (Nutta *et al,* 2002:293).

The authors point out that their results are not generalizable because of the small number of case studies in the qualitative investigations. They also discuss potentially influential factors such as attrition, IT failures and the later time in which experimental students were taught. There are, of course other potential threats to internal validity, such as the teacher inadvertently treating the experimental group differently and students having differential exposure to learning opportunities outside the experiment. However, such effects would more likely have been reflected in the immediate post-tests rather than the delayed one (unless, of course, all the extra exposure had taken place in the period between the two tests and was exclusive to the experimental group).

A serious problem of the study is the decision to set the alpha level for statistical significance at .10. This is highly unusual, it is normally set at either .01 or .05. Had this been observed the results on the delayed test (given as U=.071) would also not have been significantly different. [The U value is meaningless on its own and can only be a whole number. The alpha value, (or the two-tailed significance value) is the important

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

value and this must be less than .05 to be significant (see Gravetter & Wallnau, 2000:640-648).

The authors' discussion on possible reasons for the superior performance in the delayed tests, supported by previous findings in the area, is nevertheless interesting and worthy of further research.

6.4 Situated study design

Weasenforth *et al* (2002) conducted an interesting investigation over three semesters examining whether threaded asynchronous discussion group activities might be effective in realizing constructivist objectives (as defined by Bonk & Cunningham, 1998) in the context of postgraduate ESL (n=52). This is a descriptive (situated) study in which close observations of students' performance on frequent task-based assignments, coupled with student surveys, led to systematically documented changes in procedures and instructors' roles (interventions) in order to realize the constructivist learning objectives which formed the framework of the study.

Naturally a project of this nature cannot claim cause and effect results in terms of achievement. It does, however, provide valuable insight into the extent to which asynchronous discussions, as mediated by tutors, might promote various social and cognitive skills as well as address affective factors and motivational differences in the students. The study also contains useful models for group discussion assignments and evaluation forms.

6.5 Case study design

This study by Vincent (2001), conducted over five weeks in an elementary school (n=6), investigates the impact of multi-media software called *MicroWorlds* on the writing proficiency of children who strongly favour a visual learning style. It is an excellent example of what can be done to increase validity in a study with a very limited number of subjects and with so much scope for outcomes having been produced by elements other than the treatment. Procedures are described in great detail. Participating children were selected by rigorous selection criteria including scores from recognized (and referenced) visual, verbal and spatial tests, interviews with children and some parents and a log of classroom observations. Five children were identified as fulfilling the criteria for selection and one child performing at the opposite end of the scale, i.e. favouring a strongly verbal style, was selected as the "negative case". Assessment of the narrative writing tasks during the treatment period followed the standardized criteria devised by the Victorian Board of Studies.

The children took part in three different writing activities: (1) writing without a visual input in which they wrote by hand with the final version produced on a word processor; (2) setting up a writing task with a drama stimulus which included some use of graphics and animation with children given the choice to present their work either on paper or as a *MicroWorlds* presentation; and (3) a task in which it was compulsory to work with *MicroWorlds* software. In addition to the formal assessments of the tasks, the teacher also kept an observational journal yielding interesting information on the differences in

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

attitude towards the tasks between the children. The design is somewhat similar to 6.2 above; an added problem here, however, is the cumulative learning effect (a standard threat to within-subject designs). While the researcher does not consider this in the discussion of the results which supported activity (3), only very careful and tentative conclusions are put forward. Studies such as these are important starting points for larger experimental investigations.

7 Valid meta-analysis procedures

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 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 when 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.

Four studies in our sample can be singled out as good-practice models for meta-analyses: Bangert-Drowns, 1993, Soe *et al*. 2000, Blok *et al*. 2002, Torgerson & Elbourne, 2002. All of these concentrate their efforts around one particular aspect of learning (writing, reading, reading and spelling, respectively) 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 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

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

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 (Bangert-Drowns, 1993, Torgerson & Elbourne, 2002) 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 the Bangert-Drowns and the Torgerson and Elbourne studies give a well-documented search strategy and very comprehensive and detailed inclusion criteria. While both studies concentrated on one major variable, the Bangert-Drowns study calculated effect sizes for five subcategories: *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. This study 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 author 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 students 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.

The Torgerson and Elbourne (2002) study calculates both a confidence interval and the significance for each effect size. The Bangert-Drowns (1993) study 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 Major findings

8.1 Meta-analyses

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.

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

However, the inclusion of narrative reviews and studies carried out in L1 learning environments helps to clarify the picture somewhat. Our detailed review of studies concerned with L2 writing adds further important information.

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 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.

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 research methods in some of the primary studies 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. Our study on L2 writing lends strong support to this assertion.

8.2 Narrative reviews

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. 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

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

studies. It is interesting to note that the authors here also express concerns related to the quality of research designs and validity of findings.

8.3 L2 writing review

This synthesis of research results supports the findings reported above, especially those related to writing fluency, improved spelling and positive perceptions and attitude. This is not surprising since some of the data was derived from the same pool. It was useful, however, to also examine this data qualitatively and in more detail than was possible in the meta-analyses and narrative reviews. The reader is reminded, however, that the observations below cannot, in their own right, be generalized with the degree of confidence attributed to the findings reported above.

While the 53 studies investigated almost equally as many variables, the most frequently examined aspect was *writing quality* in various forms. Findings of interest were: a positive impact on revision strategies; a reduction in writing apprehension; higher ratings for logical thinking; the ability to switch between formal and informal language; a move from knowledge telling to knowledge building; and better awareness of the audience.

It was interesting to note that student perceptions were positively correlated with (1) their perception of the task; (2) their attitudes towards ICT; (3) the robustness of the resource; and (4) the degree of project integration. There was some indication that CALL had positive effects on motivation; computer literacy; target culture awareness; reading and listening skills; classroom climate; comfort; and participation. A trade-off effect between linguistic complexity and accuracy was observed. While spelling was reported to have been significantly improved, grammar was not.

8.4 A word of caution

It is advisable to keep in mind the authors' caution in terms of interpreting results in most studies in our sample. Interestingly, 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 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

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

means of complementary research designs generating sound qualitative and quantitative data may produce trends – some of which we have observed here – that may be interpreted with the type of confidence levels sorely needed in this controversial area.

9 Conclusions

We are beginning to see enough data in CALL that suggest positive effects on spelling, reading and writing. There is also a substantial body of data that indicates that student perceptions of CALL are on the whole positive, provided technologies are stable and well supported. On the negative side there are still concerns about technical difficulties interfering with the learning process; older students not feeling comfortable with computers; younger students not possessing the necessary metaskills for coping effectively in these challenging environments; training needs in computer literacy for both students and teachers; problems with group dynamics; and time constraints (Felix, 2004). It is sometimes forgotten that some or all of these are likely to have an influence on research results.

Now that we can examine clusters of research investigating the same or similar variables in a variety of ways, we are in a better position to ascertain how a specific item of ICT might impact a specific environment, outcome or process. A lot remains to be done, though. We need to build on existing knowledge, re-investigate established findings in different settings, replicate excellent studies using more subjects, and design sound new projects in areas and languages that have not yet been included. A design that is vastly underused in CALL research but highly recommended for longer term studies is time-series analysis in which the same group of students is involved in the experimental and control treatment for a certain amount of time and then switched – more than once if possible (see Felix, 2006b; Warschauer, 1996). Many more studies are needed in the school environment and in the Vocational Education and Training (VET) sector. We also recommend inclusion of delayed tests in designs where achievement tests are carried out, fully recognizing, of course, that it is sometimes difficult to get access to the same students.

Our systematic look at the research – starting with Dunkel (1991), examining recent work (Felix, 2005a,b) and finally filling in the gaps by investigating the specific variable L2 writing – (Felix, 2006a) 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. Further qualitative and discursive syntheses of a body of research investigating similar variables related to one larger issue such as our study on *writing*, for

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

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. It is interesting to note that the *Language Learning Journal* has dedicated an entire future issue to such work.

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.

Acknowledgements

An earlier version of this paper was delivered at the CercleS-Konferenz in Frankfurt (Oder) in 2006 and published in the conference proceedings. It is reprinted here with the permission of the convenor. Parts of this paper have previously been published in Felix (2005a; 2005b). The title has been adapted from Wigner, E. (1960) 'The Unreasonable Effectiveness of Mathematics in the Natural Sciences,' *Communications in Pure and Applied Mathematics,* **13** (1), 1-14.

References

Al-Seghayer, K. (2001) The Effect of Multimedia Annotation Modes on L2 Vocabulary Acquisition: A Comparative Study. *Language Learning & Technology*, **5**(1): 202-232.

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.

Basena, D. and Jamieson, J. (1996) CALL Research in Second Language Learning: 1990-1994. *CAELL*, **7**(1/2): 14-22.

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.

Bonk, C. and Cunningham, D. (1998) Searching for learner-centered, constructivist, and sociocultural components of collaborative educational learning tools. In: Bonk, C. and King, K. (eds.) *Electronic collaborators*. Mahwah, NJ: Lawrence Erlbaum, 25-50.

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.

Chikamatsu, N. (2003) The effects of computer use on L2 Japanese writing. *Foreign Language Annals,* **36**(1): 114-127.

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.

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

Collentine, J. (2000) Insights into the Construction of Grammatical Knowledge Provided by User-Behavior Tracking Technologies. *Language Learning & Technology*, **3** (2). http://llt.msu.edu/vol3num2/collentine/

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

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.

Denonn, L. and Egner, R. (eds) (1992) *The Basic Writings of Bertrand Russell*. New York: Routledge

Dolle, J. and Enjelvin, G. (2004) Investigating 'VLE-ffectiveness' in Languages. *Computer Assisted Language Learning*, **16**(5): 469-490.

Dunkel, P. (ed.) (1991) *Computer-Assisted Language Learning and Testing*. New York: Newbury House.

Felix, U. (2001) A multivariate analysis of students' experience of Web-based learning. *Australian Journal of Educational Technology*, **17**(1): 21-36.

Felix, U. (2004) A Multivariate Analysis of Secondary Students' Experience of Web-Based Language Learning. *ReCALL*, **16**(1): 129-141.

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.

Felix, U. (2006a) Analyzing the impact of educational technologies on L2 writing. *Paper delivered at EUROCALL*, Granada, 2006.

Felix, U. (2006b) *Accelerative Learning: Wonder method or pseudo-scientific gobbledygook*. Melbourne: CAE Press.

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

Gravetter, F. J. and Wallnau, L. B. (2000) *Statistics for the behavioral sciences* (5th ed.).London: Wadsworth/Thomson Learning.

Hoven, D. (2004) Methods in our madness? Re-examining the methodological frameworks of our field. Paper presented at the *Eleventh International CALL Conference,* University of Antwerp.

Hubbard, P. (2004) Some Subject, Treatment, and Data Collection Trends in Current CALL Research. Paper presented at the *Eleventh International CALL Conference*, University of Antwerp.

Jamieson, J. and Chapelle, C. (2004) Issues in CALL Evaluation. Paper presented at the *Eleventh International CALL Conference*, University of Antwerp.

Kulik, J. A. (2003) *Effects of Using Instructional Technology in Colleges and Universities: What Controlled Evaluation Studies Say* (No. P10446.003). Arlington, Virginia.: SRI International.

Levy, M. (2004) Interpretations of Context in CALL Research: The Goals, the Data and the Methods. Paper presented at the *Eleventh International CALL Conference,* University of Antwerp.

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.

Myers, M. J. (2000) Voice Recognition Software and a Hand-Held Translation Machine for Second-Language Learning. *Computer Assisted Language Learning*, **13**(1): 29-41.

Nunan, D. (1992) *Research methods in language learning*. Cambridge; New York, NY, USA: Cambridge University Press.

Nutta, J. W., Feyten, C. M., Norwood, A. L., Meros, J. N., Yoshii, M. and Ducher, J. (2002) Exploring new frontiers: What do computers contribute to teaching foreign languages in elementary school? *Foreign Language Annals*, **35**(3): 293-306.

Phipps, R. and Merisotis, J. (1999) What's the Difference? A Review of Contemporary Research on the Effectiveness of Distance Learning in Higher Education. http://www.ihep.com/Pubs/PDF/Difference.pdf

Robson, C. (2002) *Real world research: a resource for social scientists and practitioner-researchers* (2nd ed.) Oxford: Blackwell.

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

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

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

Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002) *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Singleton Jr., R. A., Straits, B. C. and Straits, M. M. (1993) *Approaches to Social Research* (2nd ed.). New York: Oxford University Press.

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

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

Suh, J.S. (2002) Effectiveness of Call Writing Instruction: The Voices of Korean EFL Learners. *Foreign Language Annals*, **35** (6): 669-679.

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

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.

Usability first (2002) Usability Glossary. http://www.usabilityfirst.com/glossary/main.cgi?function=search_page

Vincent, J. (2001) The role of visually rich technology in facilitating children's writing. *Journal of Computer Assisted Learning*, **17**(3): 242-250.

Warschauer, M. (1996) Computer-mediated Collaborative Learning: Theory and Practice. *Modern Language Journal*, **18**(4): 470-481.

Weasenforth, D., Biesenbach-Lucas, S. and Meloni, C. (2002) Realizing Constructivist Objectives Through Collaborative Technologies: Threaded Discussions. *Language Learning & Technology*, **6**(3): 58-86.

Yeh, S.-W. and Lehman, J. D. (2001) Effects of Learner Control and Learning Strategies on English as a Foreign Language (EFL) Learning from Interactive Hypermedia Lessons. *Journal of Educational Multimedia and Hypermedia*, **10**(2): 141-159.

Zhao, Y. (2001) Recent developments in Technology and language learning: A literature review and meta-analysis. http://ott.educ.msu.edu/elanguage/about/literature.asp

Appendix. Definitions of Research Designs used in the Studies under Investigation

PRE-EXPERIMENTAL DESIGN

May have pre- and posttreatment tests, but lacks a control group. (Nunan, 1992:41)

QUASI-EXPERIMENTAL DESIGN

Has both pre and posttests and experimental and control groups, but no random assignment of subjects. (Nunan, 1992:41)

EXPERIMENTAL DESIGN

Has both pre and posttests, experimental and control groups, and random assignment of subjects. (Nunan, 1992:41)

NONEXPERIMENTAL DESIGN

Refers to situations in which a presumed cause and effect are identified and measured but in

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

which other structural features of experiments, such as random assignment, pretests and control groups are missing. Instead reliance is placed on measuring alternative explanations individually and then statistically controlling for them.

ONE-GROUP PRETEST-POSTTEST DESIGN
A single pre-test observation is taken on a group of respondents, treatment then occurs, and a single posttest observation on the same measure follows (Shadish *et al.,* 2002:108)

NONEQUIVALENT COMPARISON GROUP DESIGN
Uses a treatment group and an untreated comparison group, with both pretest and posttest data gathered on the same units. (Shadish *et al*., 2002:136)

POSTTEST-ONLY CONTROL GROUP DESIGN
Incorporates just the basic elements of experimental design: random assignment of subjects to treatment and control groups, introduction of the independent variable to the treatment group, and a post treatment measure of the dependent variable for both groups. (Singleton Jr. *et al*., 1993:222)

PRETEST-POSTTEST CONTROL GROUP DESIGN
A design which measures the experimental group before and after the experimental treatment. A control group is measured at the same time, but does not receive the experimental treatment.

WITHIN-SUBJECTS
A study designed to make a comparison of 2 or more treatments and that compares them by having each user try each treatment, measuring their performance for each. (Usability first, 2002)

BETWEEN-SUBJECTS
A study designed to make a comparison of 2 or more treatments and that compares them by having one set of users try one treatment and another set of users try another treatment, measuring their performance for each. (Usability first, 2002)

FACTORIAL EXPERIMENTAL DESIGN
A design which enables the effects of two or more independent variables to be explored jointly. (Singleton Jr. *et al*., 1993:225)

CASE STUDY
A strategy for doing research which involves an empirical investigation of a particular contemporary phenomenon within its real life context using multiple sources of evidence. (Robson, 2002:178)

CROSS-SECTIONAL SURVEY
Data on a sample or "cross section" of respondents chosen to represent a particular target population are gathered at essentially one point in time. (Singleton Jr. *et al*., 1993:254)

NONPARTICIPANT OBSERVATION
An approach to field research in which the researcher attempts to observe people without interacting with them and, typically without their knowing that they are being observed. (Singleton Jr. *et al*., 1993:520)

TIME SERIES
Refers to a large series of observations made on the same variable consecutively over time. The

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press

observations can be on the same unit or on different but similar units. (Shadish *et al*., 2002:172)

Definitions of frequently used Terms

EFFECT SIZE
The number of standard deviation units separating scores of experimental and control groups. Values above 0.25 are large enough to be educationally meaningful (Kulik, 2003).

HAWTHORNE EFFECT
Refers to participants' awareness of being studied affecting their performance (Singleton Jr. *et al*., 1993:29).

PYGMALION EFFECT
Refers to teachers' expectations about student achievement becoming self-fulfilling prophecies (Shadish *et al*., 2002:78).

https://doi.org/10.1017/S0958344008000323 Published online by Cambridge University Press