English Teaching: Practice and Critique July, 2012, Volume 11, Number 2
http://education.waikato.ac.nz/research/files/etpc/files/2012v11n2art2.pdf pp. 25-43

Experimental methodology in English teaching and learning: Method features, validity issues, and embedded experimental design

JANG HO LEE
Korea Military Academy

ABSTRACT: *Experimental methods have played a significant role in the growth of English teaching and learning studies. The paper presented here outlines basic features of experimental design, including the manipulation of independent variables, the role and practicality of randomised controlled trials (RCTs) in educational research, and alternative methods and techniques available in the absence of RCTs. It further reviews validity issues inherent in conducting experimental research, in particular sources of internal and external invalidity, and how to remedy them. Along the way, the author suggests that researchers remain mindful of these threats, and calls for the replication of studies across different research contexts with the purposes of the cross-validation and generalisation of findings. The remainder of the paper concludes with suggestions on how to develop a more embedded and sophisticated experimental design in light of the current literature of mixed methodology (Creswell & Plano Clark, 2011), in order to have more explanatory power and compensate for the weaknesses associated with the experimental method. Throughout the paper, the author illustrates the points with examples relevant to English teaching and learning research.*

KEYWORDS: Embedded experiment, English teaching research, experimental design, internal validity, mixed methodology, quasi-experiment, randomised controlled trials.

INTRODUCTION

One of the major tasks of English teaching/learning (ETL) research is to explore complex processes and relationships among variables such as teachers' language input, various types of instructional programs, learners' literacy and attitudes towards learning environments, to name a few. All of us, as English teaching practitioners and researchers, carry hypotheses about why students behave and learn the way they do, and below are three such hypotheses for illustrative purposes pooled from some of the published examples in this field:

- University students will prefer a native speaker teacher of English in the specific areas of vocabulary, pronunciation and speaking, and a non-native speaker teacher in the areas of grammar, listening and reading (Lasagabaster & Sierra, 2005).

- As age of arrival in an English-speaking country increases, one's score on grammaticality judgment tests of English decreases (DeKeyser, 2000).

- Pre-school children who receive intensive instruction in sound categorisation will be more likely to succeed in reading and spelling than those deprived of such instruction (Bradley & Bryant, 1983).

At one level, these three hypotheses all examine the relationships among different variables (for example, age of arrival and grammatical competence), but require different research methods to be effectively addressed. For example, the first one would be best answered if we administered a questionnaire survey or conducted participant interviews, as the hypothesis is concerned with learners' beliefs or thoughts. In the case of the second illustration, the hypothesis lends itself to a correlation approach, by drawing on participants' background information (that is, age of arrival) and measuring their grammar competence via a language test, and then looking for a possible correlation between them. The final one illustrates the case in which experimental design is the most appropriate. We would divide our participants into different groups and put them into different learning conditions, and consequently compare their performance (reading and spelling skills in this case), which would be held to be the outcome of treatment effects.

The concept of experimental design, as illustrated by the final example, is the focus of the present paper. Perhaps an appropriate place to start would be with the fundamental question: Why should ETL research give any consideration at all to experimental method and design? In order to answer this question, let us first examine some statements from authors in relevant fields:

> Many people assume that the most appropriate way to resolve a question about language learning or teaching is to conduct an experiment (Brown & Rodgers, 2002, p. 195).
>
> It is the best method...of establishing cause–effect relationships and evaluating educational innovations (Dörnyei, 2007, p. 120).
>
> In experimental research on language and education an attempt is made to build theories which explain the mental processes behind language and literacy learning, the individual differences that go along with these processes and the outcomes of differential treatments meant to stimulate such processes (Verhoeven, 1997, p. 79).

Of course these statements only partially describe diverse aspects of the experimental method, but they will suffice to illustrate the point that experimental design is a crucial constituent in ETL research, which enables us to go beyond descriptive research and to seek for explanations behind the causes of students' behaviours and learning progress.

This paper will discuss basic components of experimental design, several validity issues and concerns, and methodological techniques to enhance experimental design in the context of ETL research. The primary purpose is to provide readers with sufficient background to grasp and critically evaluate examples of experimental design that they find in the literature and to help them to identify proper research designs they may select in addressing their own research questions. A major contribution of this paper is that it attempts to discuss various issues regarding experimental design in the context of English teaching and learning, considering the fact that most previous papers and books have discussed this theme in a rather broad social science or psychology domain. It will further draw on some of the recent innovations in the literature dealing with mixed methods, and consider various configurations of experimental design incorporating qualitative research elements.

EXPERIMENTAL DESIGN, RANDOMISED CONTROLLED TRIALS, AND ALTERNATIVE TECHNIQUES

Manipulating independent variables

Experimental studies aim to investigate whether there is any treatment effect on participants' behaviours or their internal processes. This entails experimental manoeuvring or artificially manipulating learning situations. A typical treatment effect in the ETL context is a teaching intervention (ranging from a classroom technique to an instructional program), through the administration of which we examine whether the intervention results in better learning. In the experimental enterprise (and education research in particular), the teaching intervention or treatment under investigation (for example, a vocabulary teaching method) is then called the "independent variable", and a set of different interventions or treatments examined in the study (for example, different techniques of teaching vocabulary) can be construed as "[different] *levels* of the independent variable" (Hinkle, Wiersma & Jurs, 2003, p. 8). On the other hand, the "dependent variable" is the one held to be the outcome of such a manipulation (Davis & Bremner, 2006).

In education research, independent variables are generally manipulated in three different ways (Johnson & Christensen, 2010). The first is "by a presence or absence technique" (p. 286), through which an experimental group receives some intervention, whereas the control group does not (see the review on Brett, Rothlein & Hurley, 1996). The second way is to manipulate the *amount* of administration of a certain intervention (Hulstijn, 1997). For example, one might presume that providing feedback to learners' writing twice is more effective than doing so once. Then one could manipulate the amount of feedback given to students by having one group which receives feedback twice, another group being given one feedback session, and the third group being deprived of any feedback. A third way is to manipulate the independent variable by varying the type of instructional techniques used. For example, Ramachandran and Rahim (2004) compared the relative effects of providing meanings of target English lexical items in the participants' first language (L1) and providing meanings of these items in English on English learners' vocabulary acquisition. English and L1 input of vocabulary thus represented two different types of teaching techniques in this study. It is noteworthy that several combinations of the above are implemented in ETL research (for example, three treatment groups respectively receiving different teaching innovations in addition to a control group with the lack thereof).

The role of randomised controlled trials

Having described the ways in which an independent variable can be manipulated, experimental studies should almost always involve two groups at least, with one being a treatment group whose members are exposed to the intervention and the other being a control or comparison group whose members are not exposed to such. However, for ethical reasons, the control group generally receives some sort of instruction – one that is not related to the target learning elements of the study (for example, Takimoto, 2008) or has regular class hours as outlined in their course syllabus (for example, VanPatten & Cadierno, 1993) during the experimentation. A true control condition without *any* instructional input or learning opportunity is rare in the field of education

research. In allocating participants to the different groups or conditions, "randomised controlled trials" (RCTs) have been labelled as the "gold standard" in measuring "what works or not" in practice and evaluating the efficacy of interventions (Delandshere, 2004). It is also viewed as "one of the hallmarks of experimental research" (Mackey & Gass, 2005, p. 146).

The primary function of RCTs is that they provide "a good chance that for each person who will respond to being an experimental subject in a particular way in one group, there will be a matching person in the other group" (Gomm, 2004, p. 26). Another way to say this would be that RCTs maximise the possibility that individual differences at the outset of a study, which are deemed extraneous factors, are "cancelled out" through the process of random allocation to either a treatment or a control group (Torgerson & Torgerson, 2001). Regarding the results of studies in which RCTs are not adopted, it has been suggested that one cannot ensure that the differences between two groups after the intervention were solely due to the effect of the intervention (Bryman & Cramer, 1994). From the validity perspective, "the presence of a control group" and RCTs "enable us to eliminate...rival explanations [of a causal finding]," and lead us to the conclusion that the intervention, other than any other factors, had the major effect on the results (Bryman, 2004, p. 35). However, it should be remembered that RCTs always presuppose a large sample size (Babbie, 2001); otherwise, the employment of RCTs cannot guarantee that "all other things" except the variables under investigation are equalised across the experimental and control groups (Torgerson & Torgerson, 2001). It is noteworthy that ETL studies are often subject to this methodological criticism, since many of them, though having administered RCTs laboriously, have fewer than 30 participants per group; thus they do not take full advantage of RCTs.

The question of whether a study has employed RCTs or not has become a criterion of good research in the field of education, and "many government requests for research proposals in education...explicitly require the use of randomised controlled trials or quasi-experimental designs but only when randomisation is not possible" (Delandshere, 2004, p. 240). The following question then is: Is it really practicable to use RCTs in the field of education research?

In general, RCTs in education are not feasible, particularly on an individual level (that is, randomly assigning each individual student to either an experimental condition or the control condition) (Moore, Graham & Diamond, 2003; Torgerson & Torgerson, 2001). To be more specific, it is difficult to employ RCTs within one school or institution, and randomly divide students in the same school into intervention and control groups, due to practical constraints. This does not mean that RCTs in other fields are always feasible or that RCTs in education cannot take place. However, the contexts of "school" and "classroom" make it rather difficult to employ individual-level RCTs under most educational circumstances. In this case, cluster sampling in which participants are randomised by group to different conditions (in education research, thus groupings of students as class or school can be the sampling unit) can be employed (Torgerson & Torgerson, 2001). A further concern arising here is the differences between classes or schools. It is very unlikely that sampled classes or schools will be comparable in terms of class size, qualifications of teachers or the proportion of genders, to name a few. These differences would unfortunately reduce

validity of the study and make it difficult to determine causal connections between an intervention and its effects on participants.

Quasi-experimental design and other methodological techniques

In the face of the reality that RCTs are often not practical in classroom contexts, researchers need to turn to quasi-experimental designs, in which intact groups are drawn on and one endeavours to equate one group with the other to the greatest possible extent. According to Fife-Schaw (2006), quasi-experiments, however, "should not be seen...as always inferior to true experiments...[they are sometimes] the next logical step in a long research process" in which research findings from laboratory experiments are tested in more practical or real life situations (p. 92). Dörnyei (2007) concurs with Fife-Schaw, proposing that "properly designed and executed quasi-experimental studies yield scientifically credible results" (p. 118).

One of the most frequent techniques used in employing the quasi-experimental design is to control for any pre-existing differences between intact groups, which are likely to be related to the outcome (dependent) variable. This is usually done in ETL research by giving some sort of a pre-test in order to statistically adjust the post-test scores between groups which may have been influenced by the initial differences between them (Dörnyei, 2007). This procedure is called Analysis of Covariance (ANCOVA), a statistical procedure that partials out the effect of a pre-existing difference (if a researcher can notice and measure this before the experimental manipulation) between the groups in order to provide a more accurate value of the *F*-ratio (Field, 2009). The statistically controlled variable in the ANCOVA procedure is called the *covariate*. For example, Lim and Shen (2006), in which the effect of Computer Assisted Language Learning (CALL) is compared with that of traditional reading classes in enhancing reading comprehension for Korean heritage learners of English at the university level, gave a reading comprehension test at the beginning (pre-test) and end (post-test) of the semester, with the pre-test being used for statistically controlling for the pre-study differences between the two groups. Thus, this method in some sense compensates for the lack of RCTs by statistically equating one group with the other in terms of a variable that is held to affect the dependent variable. Table 1 provides examples of quasi-ETL experimental studies as well as the designs and techniques they employed to compensate for the weaknesses inherent in these designs.

The matched-pairs design (Mitchell & Jolley, 2010, p. 466) offers an effective (but laborious) alternative to RCTs. The basic idea here is to identify a pair of participants who are similar to each other on several variables and distribute them to experimental and control conditions. These variables may be some important biographical variables (for example, sex, social economic status) or those suspicious ones that are likely to affect the dependent variable (for example, IQ, first language proficiency). Of course, this is easier said than done, as in reality it is a tremendous task to identify even one pair which is comparable in many aspects. As Cohen, Manion and Morrison (2011) rightly point out, we thus need to strike a balance between setting up too many variables (as a result, it would be impossible to draw on a sufficient number to sample) and too few variables (the result would be a collection of pairs which bear no resemblance) in compiling the list for pair matching.

Investigator(s)	Object of the study	Participants and designs	Strategies employed to compensate for the 'quasi' nature of the design
Brett, Rothlein, & Hurley (1996)	Examining the effects of 1) listening to stories with a brief explanation of the unfamiliar target words, 2) listening to stories without explanation, and 3) no exposure to stories on children's vocabulary acquisition	• 175 fourth-grade intact groups from six classes, respectively distributed to three conditions • Pre-test, post-test, and delayed post-test • Teachers' story reading of two books over a period of five school days for each book (except for control group)	• Intact classes were randomly distributed to each condition (one may thus suggest that RCTs were implemented on a *class* level). • The multivariate ANOVA with time as the repeated measurement was used, taking the pre-difference between the groups into statistical account.
Foorman, Francis, Fletcher, Schatschneider, & Mehta (1998)	Comparing the impacts of different types of classroom literacy instruction, which differ in the level of directness of instruction in alphabetic coding on a group of young learners in terms of reading and reading-related skills	• 285 first- and second-grade students receiving Title 1 services, with the experimental and control groups being comprised based on the willingness of the principals and teachers to participate • Repeatedly measured on reading-related skills four times a year • Reading and intelligence test at the end of the year	• The experimental and control groups were compared on word-reading and phonological awareness initially (to measure base- line differences), which are held to affect reading skill (target variable). • The implementation of intervention was monitored in order to ensure that the sampled teachers adhered to the target program.
Wilkinson & Patty (1993)	Estimating the effects of sentence combining practices on young learners' general reading comprehension and the increase in the level of their awareness of inter/intra-sentential cohesion	• Two intact heterogeneous, fourth-grade classes, with one experimental group (n = 33), and the control group (n = 32). • Pre- and post-test design • Control group received placebo treatment	• Several pre-tests were administered to ensure that the two groups were comparable on previous reading achievement and/or general intelligence. • Covariates (verbal and nonverbal I.Q.) were included in statistical procedures to minimize initial group differences.
Xanthou (2011)	Examining, in the context of science lessons at the primary level, whether instruction in the target language is more beneficial in acquiring vocabulary and content knowledge than learners' L1 instruction	• 77 intact sixth grade learners in Cyprus learning English as a foreign language (English = target language, Cypriot Greek = L1) • Pre- and post-test design • Two experiments in the same design	• Pre-tests on content and vocabulary were administered in order to 1) make sure that the two groups were comparable at the outset, and 2) measure subjects' growth. • The author replicated the experiment within her study, with an aim to lending more weight to the validity of the study.

Table 1. Examples of ETL quasi-experimental studies, and their designs and strategies employed to compensate for their weaknesses

Lastly, a repeated measures design provides another route to deal with the problem related to the unavailability of RCTs. In such a design, all participants are exposed to all treatments, and thus each participant becomes his or her own control. For example, let's say that one has 30 participants and three interventions (A, B, C) of interest. The researcher divides the participants into three groups, each consisting of ten. Each group takes different sets of interventions as per the following:

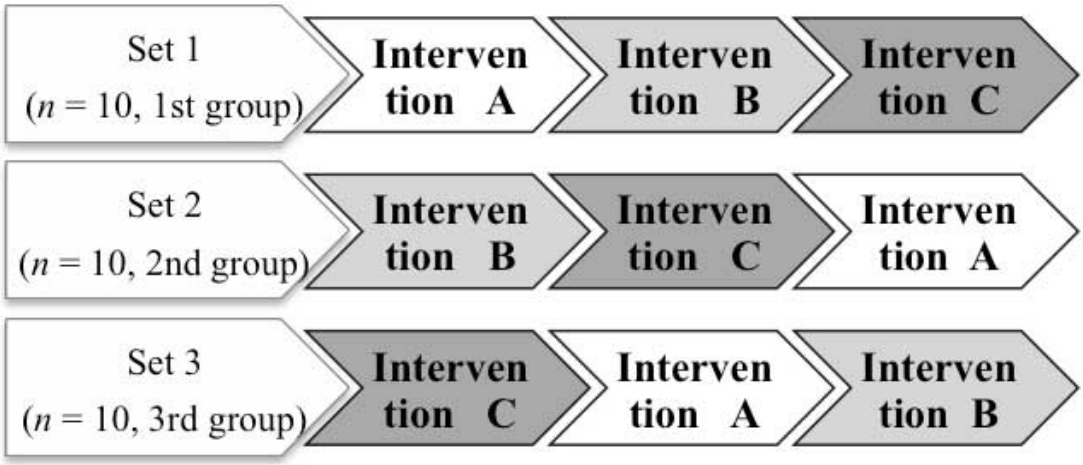

Figure 1. Ordering of three groups in a three-intervention experiment

As can be seen in Figure 1, each group is exposed to all the interventions but in a different order. This design has a methodological benefit over other experimental designs in that it "is able to factor out some of the variation that occurs within individuals because it looks at the same individuals with at least two different measures" (Larson-Hall, 2010, p. 324). It is also a more ethical one than a standard format of experimental design, in that every participant derives benefits from being exposed to the intervention, rather than being restricted to a control condition. However, as Johnson and Christensen (2010) note, researchers should plan ahead against the possible order effect – the effect resulting from a particular order in which one receives a series of interventions. It is also noteworthy that this design is rather vulnerable to some threats to internal validity, in particular those related to *maturation* and *history* (see *Threats to internal validity* below for details).

This section has considered the effectiveness and limitation of RCTs in experimental design, along with alternative approaches to examining causality in ETL research. As others (for example, Gorard, 2003; Moore et al., 2003; Torgerson & Torgerson, 2003) have proposed, RCTs can be a very powerful tool in the determination of causality when preconditions for RCTs are established. The problem is that they are frequently not feasible due to the unique features of educational contexts, and thus we need to opt for quasi-experimental designs on most occasions. However, as McDonough and McDonough (1997) argue, quasi-experiments without the implementation of RCTs may still be useful in helping one to examine questions emerging from one's teaching experience, so long as they are carefully designed and conducted. The question raising its head here is then: What are the criteria for evaluating a piece of ETL research using the experimental method? The answer to this question lies in the notion of validity, to which we turn next.

VALIDITY ISSUES AND EXPERIMENTATION IN LANGUAGE CLASSROOMS

Carrying out experiments in language classrooms is not without problems and limits, and English teachers and researchers who embark on their own research with this method and/or who need to gain knowledge from literature of this kind should be acquainted with overall design features and aspects of this particular method as well as relevant procedures. Among the various aspects of experiments is the validity of the experiments which is one of the most essential issues for one to critically evaluate the quality of a piece of research, and consequently tease out the implications from the findings therein. Validity takes an important position in experimentation, because it is concerned with "the truth of the causality" and it is "a basic tenet of experimental method" (Davis & Bremner, 2006, p. 73).

Before we move onto validity issues in depth, there are some characteristics of language classrooms, which are concerned with validity in ETL research. As van Lier (1988) and other educational research texts (for example, Cohen et al., 2011) rightly point out, the classroom, like any other social milieu, does not function in a vacuum. As van Lier (1988) puts it, "As yet we know too little about all the variables that play a role in all the classrooms to be able to make rash recommendations about methods of teaching and ways of learning" (p. 7). Although ETL and relevant fields of study such as *Applied Linguistics* and *Second Language Acquisition* have expanded their research coverage over the last two decades in exploring classroom-related variables, the above point is still valid today. In addition, the language classroom, like other educational settings, is a difficult place in which researchers and teachers do not always have the power to manipulate variables and contexts, not to mention the difficulty attached to the implementation of RCTs and ethical and methodological issues involved in dealing with human participants. These characteristics often generate some threats to the validity of studies, which I will discuss below, using illustrations relevant to ETL interests.

Threats to internal validity

Internal validity is the extent to which the observations and interpretations of a causal connection drawn from research hold true (Bryman, 2004). To put it differently, it is the degree to which the experimental arrangements of a study rein in spurious variables that might compromise the integrity of the causal link between the independent and dependent variables (Borg, Gall & Gall, 1993). In *Quasi-Experimentation*, Cook and Campbell (1979) identified different types of threats that lurk in the conducting and interpreting of one's own research and this section will review some of those most relevant to ETL research:

> *History:* This threat relates to historical events other than experimental manoeuvring that take place during the course of a study (particularly between the pre- and post-tests). This effect may be more threatening to research on participants who learn English as an additional language (EAL) or who are mature and experienced enough to bring their own learning strategies into their English learning, as it is possible that they would review what they learn in researchers' manipulated environments outside the research contexts.

Maturation: During the course of the treatment, physical or cognitive changes may occur to participants. The maturation effect highlights the importance of having control groups in experimental designs, the lack of which will elicit such criticism that any change in the participants' performance may be attributed to their maturation, rather than the treatment effect (that is, maturation will presumably affect both groups equally). This issue is particularly relevant to research on early childhood, which reflects remarkable physical, cognitive and emotional development even over a short period of time.

Selection: Researchers need to look out for any pre-existing differences between the experimental and control groups (which may have resulted from a non-random sampling procedure), otherwise differences between them in any measures after the experimental manipulation cannot be attributed to a treatment effect.

Selection-maturation interaction: Thye (2007) outlines this threat clearly, suggesting that selection biases and maturation processes may interact with each other in some unexpected manners, and bring about a combined effect. He gives an illustration in which a considerable proportion of gifted children are assigned to the treatment condition, and their improvement in scores compared to normal ones in the control group is caused by an interactional effect between maturation and their talent, which boosts their learning processes, rather than being due to the treatment.

Testing: This validity threat is evident when a particular measure is repeatedly given to the participants. In Cook and Campbell's own words (1979), "familiarity with a test can sometimes enhance performance," as participants are likely to remember the same or similar items "at later testing sessions" (p. 52). Ellis (1994) similarly notes that repeated measurements may affect "subsequent levels" in an unpredictable fashion.

Diffusion or imitation of treatments: In some educational contexts, it is possible that experimental and control groups communicate with each other, and the participants in the experimental group may reveal information about the treatment they are currently receiving. It follows then that the control group's performance on a post-test cannot be considered valid, as they are exposed to the treatment, albeit in an indirect manner. As Cook and Campbell (1979) note, this is a more relevant issue in conducting quasi-experiments (non-laboratory environment) in which participants in different conditions are usually permitted to interact with each other.

Mortality: This refers to participant dropout during the experiment. The loss of participants is indeed a serious practical concern for researchers, but it is more devastating from the validity perspective, as the loss of participants in a certain group is usually due to non-random reasons. In general, those who stay are more tolerant and motivated learners than those who leave, and this phenomenon is likely to bias research findings.

These sources of internal invalidity can be best controlled for by randomly assigning participants to different conditions. Thye (2007), for example, suggests that several effects concerning history, maturation, selection, selection-maturation interaction, and testing would influence both experimental and control groups evenly. Other less controlling designs (quasi-experimental design without RCTs) would then be more vulnerable to these threats. In the face of reality, in which RCTs are difficult to adopt, the best approach one can take is to conceive designs which guard against threats so far as circumstances permit, and carefully administer such designs. The rest of this section shall briefly review one ETL study and evaluate it from the perspective of internal validity.

An illustration

Carter, Ferzli and Wiebe (2004) conducted a quasi-experimental study comparing the effectiveness of two kinds of genre teaching methods, in teaching the genre of the laboratory report to university students enrolled in biology labs. The independent variables included a set of online instructional materials of their own creation – LabWrite –, which "is structured as a guide to the lab experience, organised as a chronological process paralleling the lab activities" (p. 400), as well as a more traditional approach, simply providing a handout describing each section of the lab report (for example, how each section should be written in a science report). The treatment group consisted of students registered for a biology course in the Spring semester, 2001, while the control group was registered in the Fall semester, 2000. The treatment and control groups were taught by the same professor (in the lecture session) and the same instructors (in the lab sessions), and the courses both groups took were basically identical in terms of the syllabi, labs and assigned reports. Students' lab reports were collected and later analysed in terms of their understanding of the scientific concepts that the lab experiments were supposed to reinforce and their ability to apply scientific reasoning, as reflected in their reports. An attitudinal questionnaire regarding their attitudes towards lab reports was also given to the participants later. The results showed that there were statistically significant advantages for LabWrite over the traditional approach for participants to learn the scientific concepts of the lab as well as increase their ability to apply the elements of scientific reasoning to lab experiments. It was further found that the treatment elicited a significantly more positive attitude towards lab reports than the traditional approach.

The study by Carter et al. (2004) is a good illustration of the advantage of (quasi-) experimental design, in which the effect of a newly designed pedagogical approach can be compared with that of an existing one. However, the findings from this study need to be read with some caution in view of the threats to internal validity discussed above. First, how can the researchers make sure that there were few differences between the experimental and control groups without having administered RCTs (selection bias)? Also, if the treatment group developed significantly in terms of the target variable, is this solely due to LabWrite (that is, they might have also done something else outside the classroom – history effect)? In fact, they administered a demographic survey to their participants in order to see whether the control and treatment groups were comparable, and they found that the control group was comprised of more advanced participants in terms of academic class and the number of science courses they had taken. Thus, the two groups were not comparable, though the treatment group eventually fared better their control counterpart, fortunately verifying the authors' hypothesis. However, the lack of a pre-test in the design still

makes it difficult to estimate the extent to which the treatment effect reported in the study was not distorted by the methodological limitations described above.

Threats to external validity

Another important validity issue is concerned with the extent to which the findings of a study can be generalised across different populations and contexts (Campbell, 1957; Cook & Campbell, 1979). The two threats that will be presented here are *interaction of selection and treatment*, and *reactive or interactive effects of testing. Interaction of selection and treatment* relates to the issue of how confident one can be that study findings are generalisable to other learner groups. For example, it is possible that learners of the same age and nationality may show different learning outcomes, depending on whether they learn English in urban or suburban areas. On the other hand, *reactive or interactive effects of testing* raises the question of whether pre-testing can sensitise subjects to the extent that their cognitive status does not resemble that of ordinary students in non-research settings. The consequences of this effect may vary: participants may attempt to work hard in post-tests to meet researchers' expectations or they may pay an undue amount of attention to a post-test as a result of being ed by the test.

Threats to external validity are obviously a tormenting issue for a researcher, as most of his or her readership would be more interested in the extent to which the findings of a study could be generalised to their own teaching contexts than what actually happened in the context of the study at hand. One obvious but laborious solution to *interaction of selection and treatment* is to replicate the study. Indeed, such an effort often reveals that one type of pedagogical method found to be effective for one population may not be so for others. Let's take the example of Eckerth (2008), who investigated the effects of a series of dyadic consciousness-raising (CR) tasks on the acquisition of grammatical elements with participants in two regular German-as-a-target-language university courses and found that two types of CR tasks (text reconstruction and text repair) were beneficial to students' learning outcomes. This finding was not confirmed in a quasi-replication study by McNicoll and Lee (2011), in which the same types of CR tasks were administered to Korean EAL learners; significant learning gains were demonstrated only as a result of the text repair tasks, with the text reconstruction tasks being found to be rather cognitively demanding for the participants in the study. That said, the study by McNicoll and Lee, along with the findings of Eckerth (2008), points to an otherwise unrevealed possible interaction between CR tasks and different groups of learners.

Some threats, for example *reactive or interactive effects of testing*, on the other hand, can be guarded against by the administration of a more sophisticated experimental design. *Solomon Four-Group Design* (Campbell & Stanley, 1963) attempts to control for the effects of testing interaction with the treatment. Although space constraints do not allow for a detailed description of this design, the basic idea is to have the four possible group combinations derived from the following two criterion dimensions: the first being the administration of a pre-test (taking part in the experiment with or without a pre-test) and the second being the condition (being exposed to an intervention or not). Campbell and Stanley point out that this design not only estimates the effects of *testing*, but also provides additional lenses through which we can examine the interactive effects between *treatment* and *testing*. The shortcoming of

this design, of course, is that it requires a greater number of participants per group to make proper statistical inferences about the population from which the participants are drawn.

Ecological validity and natural experiment

Another type of validity, which is seen as part of external validity in some literature but often treated separately, is ecological validity (Bronfenbrenner, 1976), concerning the applicability of findings from laboratory experiments (or more controlled settings) to real life pedagogy (Hulstijn, 1997). This issue is a serious one, in particular if we consider it from the vantage point of participants. For example, the controlled setting in which participants are situated can be different from their ordinary classroom environments in several aspects (for example, seating arrangement, acoustics of the room). In many cases, a researcher who provides treatment is not their ordinary English teacher; rather it could be an outsider. The intervention of a researcher's interest (usually an innovative one) could be something that participants are not likely to encounter in their real life classrooms (Schmidt, 1994) and this could have an inadvertent effect on their learning consequences. All of these characteristics may render the findings from the experiment of limited pedagogical value.

One experimental design devised to address this problem is called a "natural experiment" (Babbie, 2001) in which researchers exert little control over variables (including independent and extraneous ones), drawing on naturally occurring findings. In the context of classroom research, a natural experiment may be conducted when one can get access to participants who have already experienced a particular treatment of interest (rather than providing such a treatment after a sampling procedure), and give a test measuring the dependent variable on which they aim to estimate the effect of a treatment. Below is an example of a natural experiment study.

Illustration

Kobayashi and Rinnert (2008) were interested in investigating the effects of previous intensive preparatory training in L1 (Japanese) and/or L2 (English) essay writing (for university entrance exams) on the task response and structural features in L1 and L2 essays. The independent variable was the type of essay training the participants received: intensive writing experience in both the L1 and L2 (n = 9), intensive writing experience in only the L1 (n = 7), intensive writing experience in only the L2 (n = 7), no intensive writing experience in either language (n = 5). It is noteworthy that the researchers did not provide these different types of training; rather they sampled participants with different writing experience who met their sampling criteria. The participants were asked to write two essays, one in their L1 and the other in English. Their essays were analysed in terms of their use of discourse type and discourse markers in their essays (that is, dependent variables). Since RCTs were not administered, the authors ensured via their scores on an English proficiency test that the four groups were more or less comparable. The participants were also interviewed later regarding their composing processes and writing background. The interview data were used to cross-validate and supplement the findings from their analysis of the essays.

The study found that previous training experience had differential effects on participants' writing competence. The L1 intensive writing experience pointed to the

importance of increasing clarity in writing and establishing original thinking, whereas the L2 writing training emphasised the need to lean to one position in making an argument and to make this explicit in the introduction and end of an essay. Intensive writing experience in both languages, on the other hand, raised the learners' awareness of various aspects of L1 and L2 writing at the discourse level, resulting in the application of meta-knowledge in both types of writing.

This kind of study is deemed to be of high ecological validity in terms of the effect of the independent variable in that treatments were "naturally" given to the participants in their ordinary situations. Although one could say that the dependent variable (that is, essay prompts) was manipulated to some extent by the authors (as the essay prompts were those prepared by them, rather than naturally occurring ones), it is still a relatively less controlling one compared to post-tests administered in other (quasi) experimental studies. However, the study is subject to a number of methodological criticisms:

- The *data collection* section does not provide sufficient information about the context in which the essay writing took place (for example, was it conducted by their ordinary English teacher?).

- The study was a small-scale study (less than ten participants per group), and thus any conclusion has to be read with extreme caution.

- The authors are honest in stating that all the participants had some L1 writing instruction and experience throughout the primary and secondary levels in Japanese language classes (p. 11), but do not provide us with much explanation thereof, which would enable readers to estimate its potential effects.

- Only one composition topic was used for each language, and thus we cannot know whether the topic might have affected their composition. Indeed, they acknowledge in their conclusion that they "observed a possible topic effect" (p. 20) and this might have weakened the internal validity of the experiment. To be fair, the authors state that their study is exploratory by nature, and do not claim their study to be "an experiment" anywhere in the text (though they use phrases such as *to explore possible effects of ... on* and *how various types of ... affect L1 and L2 essays*, implying some aspect of experimentation).

These methodological limitations, in addition to the "naturalness" of the experiment, however, undermine the internal validity of the study, and thus make the study less worthwhile for those who put priority on determining causality. It should be noted that the criticism concerning internal invalidity is almost always likely to be associated with a natural experiment, no matter how well it is devised and written. This criticism may be seen as unfair somehow, as natural experiments by definition lack experimental manipulation. Instead, we need to acknowledge that it is a daunting task to create a balance between establishing any causal chain regarding language learning processes and generalising findings in view of practical constraints. That is to say, it is extremely difficult for findings to be both ecologically and internally valid. A question emerging from this discussion lies not so much in the relative necessity of natural and laboratory experiments, but in the issue of where we should draw the line

along the continuum of experimental research designs. Unfortunately, there is no hard-and-fast rule regarding this issue, and the answer to the above question may largely depend on the value one puts on causation and generalisablity of findings to real life pedagogy.

One possible way to address this difficult problem would be to conduct two experiments within a single study or across a series of studies, with one in a more controlled setting and the other in a setting which foregoes some aspects of experimental manipulation, and then to compare the results in two different contexts. It would be the aim of such a study to confirm and cross-validate its findings (and hopefully there would be some consistency in the findings or at least some distinct patterns from which researchers may deduce contextual effects). This may, of course, take more practical resources, but it would certainly put researchers in a much better position to argue for their findings with confidence. It has been found to be fruitful to pursue such cross-validation in the fields of industrial-organisational psychology and organisational behaviour, in which notable consistency has been observed in the findings of laboratory and field-based studies (Locke, 1986). Whether this similarity would be mirrored in ETL studies, unfortunately, is open to speculation. And as Hulstijn (1997) rightly points out, in future research we should conduct laboratory experiments as well as studies using natural research methods, which would shed more light on the interaction between contextual factors and language learners' acquisition processes.

EMBEDDED EXPERIMENTAL DESIGN

The final issue to be discussed is the idea of integrating qualitative research elements into experimental designs. Although some may oppose harnessing qualitative instruments concurrently with experimental methods, there are some good reasons for adopting such elements from qualitative research. A major rationale for doing so comes from the inherent weakness of the experiment – that a lack of qualitative description of the phenomenon being examined does not give us a complete picture of what happens in the causal connection between the treatment and outcome (Moore et al., 2003). In Howe's (2004) own description: "Acquiring a better understanding of causal mechanisms requires substantive knowledge of the contents and workings of the black box, something that cannot be obtained merely by employing the formal device of the randomised experiment" (p. 47).

It is now generally accepted that experimental design and RCTs can be greatly enhanced in terms of their ability to account for causal connections by integrating qualitative elements (Goldstein & Blatchford, 1998; Howe, 2004; Raudenbush, 2005). That said, the research design of our interest would be the experimental mixed methods design, in which the experiment takes the leading role, with the additional collection and analysis of qualitative data taking place before, during, or after the implementation of the experiment. Creswell and Plano Clark (2011) label this "the embedded-experiment variant" (p. 95) as one type of the embedded design, and further show that the qualitative elements of this design can offer more than what most people would assume, which will be discussed below in light of their own description and ETL contexts (pp. 92-93):

- *Develop outcome measures and intervention:* One way for this to occur would be to conduct a pilot study using qualitative instruments (for example, conducting interviews with those who are similar in profile to the prospective participants of the main study) before one attempts to devise his or her own intervention, or improve outcome measures (that is, via asking participants questions regarding the difficulty and content of the test) to the characteristics and levels of target participants.

- *Describe participants' experiences with the intervention:* We may conduct in-depth interviews with those who participate in an experiment. Johnson and Christensen (2010) suggest that parallel interviews would reveal participant views and allow us to have a better grasp of findings derived from the experiment and the meaning behind the numbers. It is recommended that interviews of this sort be conducted after the completion of the experiment, as interviews in the middle of experimental sessions may raise participants' consciousness of both intervention and target variables, and cause the Hawthorne effect, which is defined as "the tendency of human beings to temporarily improve their performance when they are aware it is being studied" (Singh, 2007, p. 67). Adopting stimulated recall methodology (see Gass & Mackey, 2000 for a full description of how to utilise this methodology in language teaching and learning research) may also be useful in this regard, asking participants to recall their thoughts while they were engaging in a particular task or taking a target treatment. In such a process, participants are often aided with audio- or video-recorded data on their on-the-spot performance at the moments of intervention sessions, in order to stimulate their memories more effectively.

- *Describe the process and treatment fidelity:* One can directly observe the situation in which an experiment is being conducted, and attempt to understand how and why a treatment works or not for one's participants in a particular context. Engaging in observation will also be informative in determining whether the treatment has been accurately and/or authentically carried out in the fashion the researcher originally intended (especially if it is being implemented by outsiders). The pitfall of doing so is that observation could be obtrusive to some participants, and is likely to exert some influence on their learning process (and thus create another extraneous variable). Not informing participants of the fact that they are involved in an experiment, however, may raise ethical concerns (Bryman, 2004), which somehow point to the conflict between participants' right to know what they are being exposed to and researchers' deliberate deceptions for the purpose of increasing validity.

- *Describe what long-term effects are experienced:* The experiment could be followed by a case-study or ethnography on a small scale (see Richards, 2003 for guides to implementing these types of methods in the ETL context) to examine the extent to which the intervention or treatment of the experiment has an enduring effect on participants' language and learning behaviours, from both psycholinguistic and sociolinguistic perspectives.

Despite the strengths of such a design, Creswell and Plano Clark (2011) note that designing and implementing an embedded study may raise some challenges for

researchers, as it not only requires expertise in quantitative and qualitative design, but also in mixed methods research, to properly execute such a design. Dörnyei (2007), while acknowledging that most researchers are inclined towards either quantitative or qualitative research, and suggest that researchers with different orientations may work in teams to overcome this challenge. More optimistically, the number of research methodology texts (for example, Greene, 2007; Hesse-Biber, 2010; Morse & Niehaus, 2009; Teddlie & Tashakkori, 2009) as well as journals (for example, *Journal of Mixed Method Research*, *International Journal of Mixed Methods in Applied Business & Policy Research*) relating to this issue is increasing, from which researchers and teachers may gain knowledge regarding how to design, implement and report on the mixed design study.

CONCLUDING REMARKS

The present paper has attempted to acquaint readers with some of the basic elements of experimental design and several validity issues in the ETL field, as well as to make some recommendations to compensate for its weaknesses. It has been suggested that experimental design is used in ETL research to assess the effects of new teaching innovations, and that RCTs are significant vehicles for empowering researchers to ensure that the independent variable is the most likely one to affect the dependent variable, by ruling out alternative explanations regarding the connection between them. In view of the reality that RCTs are not feasible in several situations in the context of classroom research, quasi-experimental designs and other methodological techniques within the experimental framework were introduced, followed by several threats to the different types of validity.

It is hoped that these pages on the list of threats to validity have not frightened readers away from using the experimental design in their own research. To the extent that researchers remain attentive to these threats, they will be more prepared to cope with them when approaching relevant literature and conducting their own research projects. Indeed, one should acknowledge the weaknesses involved in carrying out (quasi-)experiments in educational contexts, and honestly take them into account when designing, implementing and reporting on experimentally propelled projects. The study by Wilkinson and Patty (1993) is an excellent example, in which they describe potential sources of threats, which could undermine the validity of their design along with the methodological steps taken to alleviate or minimise them. It is unfortunate that we rarely find a study like this one with a detailed description of validity issues in its methodology section.

This paper went on to argue that integrating qualitative elements provides a means of compensating for the shortcomings of the experimental design. Creswell and Plano Clark's (2011) notion of the embedded experimental mixed method was outlined along with its implications for ETL research. It is expected that this design will be more widely adapted in future research, enhancing the current format of the experimental design and thus allowing us to gain a better insight into the causal mechanism at hand. It was also stated that this design would benefit greatly from the team approach, in which researchers with different research orientations could infuse their expertise to examine the target phenomenon more successfully. The present paper lastly suggests that, for novice researchers, graduate students and prospective

teacher researchers, methodology courses and research method textbooks regarding experimental design should be imbued with the notions of different configurations of the embedded experimental design, which would open up a wider avenue of research enquiry in exploring various ETL issues concerned with establishing causal relations.

REFERENCES

Babbie, E. (2001). *The practice of social research* (9th ed.). Belmont, CA: Wadsworth/ Thomson Learning.

Borg, W. R., Gall, J. P., & Gall, M. D. (1993). *Applying educational research: A practical guide* (3rd ed.). New York, NY: Longman.

Bradley, L., & Bryant, P. E. (1983). Categorising sounds and learning to read - A causal connection. *Nature*, *301*, 419-421.

Brett, A., Rothlein, L., & Hurley, M. (1996). Vocabulary acquisition from listening to stories and explanations of target words. *Elementary School Journal*, *96*(4), 415-422.

Bronfenbrenner, U. (1976). The experimental ecology of education. *Educational Researcher*, *5*(9), 5-15.

Brown, J. D., & Rodgers, T. S. (2002). *Doing second language research*. Oxford, England: Oxford University Press.

Bryman, A. (2004). *Social research methods* (2nd ed.). Oxford, England: Oxford University Press.

Bryman, A., & Cramer, D. (1994). *Quantitative data analysis for social scientists*. London, England: Routledge.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297-312.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Carter, M., Ferzli, M., & Wiebe, E. (2004). Teaching genre to English first-language adults: A study of the laboratory report. *Research in the Teaching of English*, *38*(4), 395-419.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (7th ed.). London, England: Routledge.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand McNally.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.

Davis, A., & Bremner, G. (2006). The experimental method in psychology. In G. M. Breakwell, S. Hammond, C. Fife-Schaw & J. A. Smith (Eds.), *Research methods in psychology* (3rd ed.) (pp. 64-87). London, England: Sage.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*, 499-533.

Delandshere, G. (2004). The moral, social and political responsibility of educational researchers: Resisting the current quest for certainty. *International Journal of Educational Research*, *41*(3), 237-256.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies.* Oxford, England: Oxford University Press.

Eckerth, J. (2008). Investigating consciousness-raising tasks: Pedagogically targeted and non-targeted learning gains. *International Journal of Applied Linguistics*, *18*(2), 119-145.

Ellis, L. (1994). *Research methods in the social sciences*. Madison, WI: Brown & Benchmark.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London, England: Sage.

Fife-Schaw, C. (2006). Quasi-experimental designs. In G. M. Breakwell, S. Hammond, C. Fife-Schaw & J. A. Smith (Eds.), *Research methods in psychology* (3rd ed.) (pp. 88-103). London, England: Sage.

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Education Psychology*, *90*(1), 37-55.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Goldstein, H., & Blatchford, P. (1998). Class size and educational achievement: A review of methodology with particular reference to study design. *British Educational Research Journal*, *24*(3), 255-268.

Gomm, R. (2004). *Social research methodology: A critical introduction.* Basingstoke, England: Palgrave Macmillan.

Gorard, S. (2003). *Quantitative methods in social science*. London, England: Continuum.

Greene, J. C. (2007). *Mixed methods in social inquiry*. San Francisco, CA: Jossey-Bass.

Hesse-Biber, S. (2010). *Mixed methods research: Merging theory with practice*. New York, NY: Guilford.

Hinkle, D. E., Wiersma, W., & Jurs, S.G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston, MA: Houghton Mifflin Company.

Howe, K. (2004). A critique of experimentalism. *Qualitative Inquiry*, *10*(1), 42-61.

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

Johnson, B., & Christensen, L. (2010). *Educational research: Quantitative, qualitative, and mixed approaches* (4th ed.). Thousand Oaks, CA: Sage.

Kobayashi, H., & Rinnert, C. (2008). Task response and text construction across L1 and L2 writing. *Journal of Second Language Writing*, *17*, 7-29.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York, NY: Routledge.

Lasagabaster, D., & Sierra, J. M. (2005). What do students think about the pros and cons of having a native speaker teacher? In E. Llurda (Ed.), *Non-native language teachers: Perceptions, challenges, and contributions to the profession* (pp. 217-242). New York, NY: Springer.

Lim, K-M., & Shen, H. Z. (2006). Integration of computers into an EFL reading classroom. *ReCALL*, *18*(2), 212-229.

Locke, E. A. (1986). Generalising from laboratory to field: Ecological validity or abstraction of essential elements? In E. A. Locke (Ed.), *Generalising from laboratory to field settings* (pp. 3-9). Lexington, MA: Lexington Books.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design.* Mahwah, NJ: London: Erlbaum.

McDonough, J., & McDonough, S. H. (1997). *Research methods for English Language teachers*. London, England: Arnold.

McNicoll, J., & Lee, J. H. (2011). Collaborative consciousness-raising tasks in EAL classrooms. *English Teaching: Practice and Critique*, *10*(4), 127-138.

Mitchell, M. L., & Jolley, J. M. (2010). *Research design explained* (7th ed.). Belmont, CA: Thomson Wadsworth.

Moore, L., Graham, A., & Diamond, I. (2003). On the feasibility of conducting randomised trials in education: Case study of a sex intervention. *British Educational Research Journal*, *29*(5), 673-689.

Morse, J. M., & Niehaus, L. (2009). *Mixed method design: Principles and procedures*. Walnut Creek, CA: Left Coast Press.

Ramachandran, S. D., & Rahim, H. A. (2004). Meaning recall and retention: The impact of the translation method on elementary level learners' vocabulary learning. *RELC Journal*, *35*(2), 161-178.

Raudenbush, S. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher*, *34*(5), 25-31.

Richards, K. (2003). *Qualitative inquiry in TESOL*. New York, NY: Palgrave Macmillan.

Schmidt, R. (1994). Implicit learning and the cognitive unconscious: Of artificial grammars and SLA. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 165-209). London, England: Academic Press.

Singh, K. (2007). *Quantitative social research methods*. London, England: Sage.

Takimoto, M. (2008). The effects of deductive and inductive instruction on the development of language learners' pragmatic competence. *The Modern Language Journal*, *92*(3), 369-386.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research*. Thousand Oaks, CA: Sage.

Thye, S. R. (2007). Logical and philosophical foundations of experimental research in the social sciences. In M. Webster & J. Sell (Eds.), *Laboratory experiments in the social sciences* (pp. 57-86). San Diego, CA: Elsevier.

Torgerson, C. J., & Torgerson, D. J. (2001). The need for randomised controlled trials in educational research. *British Journal of Educational Studies*, *49*(3), 316-328.

Torgerson, D. J., & Torgerson, C. J. (2003). Avoiding bias in randomised control trials in educational research. *British Journal of Educational Studies*, *51*(1), 36-45.

van Lier, L. (1988). *The classroom and the language learner*. London, England: Longman.

VanPatten, B., & Cadierno, T. (1993). Input processing and second language acquisition: A Role for Instruction. *The Modern Language Journal*, *77*(1), 45-57.

Verhoeven, L. (1997). Experimental methods in researching language and education. In N. H. Hornberger & C. Corson (Eds.), *Encyclopedia of language and education, Volume 8: Research methods in language and education* (pp. 79-87). Dordrecht, The Netherlands: Kluwer Academic.

Wilkinson, P. A., & Patty, D. (1993). The effects of sentence combining on the reading comprehension of Fourth Grade students. *Research in the Teaching of English*, *27*(1), 104-125.

Xanthou, M. (2011). The impact of CLIL on L2 vocabulary development and content knowledge. *English Teaching: Practice and Critique*, *10*(4), 116-126.

Manuscript received: March 9, 2012
Revision received: May 23, 2012
Accepted: July 7, 2012