

Can educational experiments both test a theory and inform practice?

Dr. Emma Marsden¹

Accepted for publication in the British Educational Research Journal, 2007.
Not to be cited without author's permission

University of York, UK.

Contact details:

Department of Educational Studies

University of York

York

YO10 5DD

Tel: 0113 2737129 / 01904 433335

em502@york.ac.uk

Can educational experiments both test a theory and inform practice?

Abstract

This article feeds into debate about the feasibility and usefulness of educational experiments by discussing methodological issues arising out of a study which sought causal links between teaching and learning of one aspect of French as a foreign language. The study involved two small scale experiments which tested a hypothesis regarding the learning of second language grammar within a particular theoretical, and its related pedagogical, framework (Input Processing and Processing Instruction respectively), and has been described in full elsewhere. The current article uses that example to suggest some circumstances (contextual, methodological and theoretical) within which educational experiments may be able to both test a learning theory and inform educational practice. It is argued that despite the complexities and limits of small scale educational experiments, an experimental design which combined a range of methods was able to generate new and useful (in a range of senses) substantive knowledge.

Can educational experiments both test a theory and inform practice?

Introduction: Rationale for a methodological evaluation of an experiment which aimed both to test a learning theory and inform practice

There have been many contributions to debates about whether experimental approaches are feasible and/or appropriate for educational research (e.g. Hargreaves, 1996; Atkinson, 2000; Oakley, 2001; Cook & Payne, 2002; Gorard, 2002c; Thomas & Pring, 2004; Lather, 2004). In fact, very few researchers actually deny that it may be possible to accumulate knowledge about social life which is indicative of causal explanations e.g. Hammersley (1997 & 2001) and Pring (2000) (though see Morrison, 2001; Bruner, 1990). However, there is considerable controversy about the nature of the knowledge generated by educational experiments and its usefulness to practitioner and/or research communities (Bassey, 1995; Elliot, 2001; Hammersley, 2003). Whether and how large-scale experimental designs can inform *practice* has been discussed in some depth (e.g. Torgerson & Torgerson, 2001 & 2003; Moore *et al.*, 2003; Slavin, 2002), and roles have been suggested for the small scale experiment within grander research designs which, in turn, intend to inform practice and policy e.g. design studies (e.g. Shavelson *et al.* 2003; Gorard *et al.* 2004), Randomised Control Trials (e.g. Slavin, 2002; Torgerson & Torgerson, 2003) and systematic reviews (e.g. Evans & Benfield, 2001; Andrews *et al.*, 2004; REEL, 2004). However, there has been little focus on whether and how an educational experiment can test a specific learning hypothesis *and* inform practice. This article sets out some specific suggestions in this regard.

Why might it be desirable for one study to both test a learning theory and inform practice?

It has been argued extensively that one of the purposes of educational research is to inform educational practice at some level, in some way, regardless of the methodological or epistemological perspectives of the researchers (e.g. Stenhouse, 1980; Hargreaves, 1996; Hiebert *et al.*, 2002; Thomas & Pring, 2004). With regards to educational research informing theory, arguments have also been made that theory development is desirable in order to search for improvements to educational practice which can transcend idiosyncratic classrooms (Bassey 1984, Foster 1999), regardless of the research method in focus. See, for example, Brown (1992) and Shavelson *et al.* (2003) regarding design experiments; Stenhouse (1980) and Hiebert *et al.* (2002) regarding collaborative action research approaches; Moore *et al.* (2003) regarding large scale Randomised Control Trials. Hammersley (2003) proposes “It would not be logically contradictory, or even obviously irrational, to argue that the best way for a researcher to contribute to the improvement of education would be to build sociological and psychological knowledge in this field” (p.12). In brief, it seems that finding out whether a theory of learning and/or teaching (regardless of how the term ‘theory’ is used), is upheld or refuted in the messy environment of teaching practice, is a desirable way of refining theory and informing practice.

However, several researchers have suggested that in fact genuine educational research does not, or even cannot, test learning theories. Bassey (1984) drew distinctions between “disciplinary research” and “pedagogic research” in education, suggesting that the former is characterised by the methodology, concepts and aims of a parent discipline (e.g. psychology, sociology). Bassey (1995) argued that disciplinary research was rejected by practitioners as being of little or no use to their practice

because it “aims critically to inform understandings of social [or psychological, etc] phenomena in educational settings” (pp. 42–45) whereas educational research aims to “critically to inform educational judgments and decisions in order to improve educational action” (p. 39). Elliott (1991) drew similar distinctions between the ‘production of knowledge’ (non teacher research) and the ‘improvement of practice’ (action research)². Others argue that most educational research designs and their reports do not lead to theories being falsified (Gorard with Taylor, 2004: 144), that the theories are expressed so that they cannot be falsified, or that genuine educational research cannot inform learning theories in the same way that social science disciplines build and test theories due to the nature of education itself as a distinctly human and moral enterprise (e.g. Chambers 1992; Pring 2000; Thomas 2002 & 2005).

Indeed, the emergent interest in design studies (also termed design experiments) has, in part been motivated by a desire to negate the necessity of a pre-established theory so that educational research may inform practice in situations which are “often poorly understood theoretically” (Gorard *et al.* 2004: 586, and Kelly & Lesh 2002). The suggestion that theory is not a pre-requisite for useful educational research is also supported elsewhere (e.g. Tymms & Fitz-Gibbon 2002; Thomas 2002 & 2005), and the arguments in the current paper do not preclude that suggestion. The relevant point here is that the aims and methods of design studies (e.g. a flexible intervention protocol) mean they do not have an obvious capacity for testing (i.e. falsifying), in the classroom, focused hypotheses about learning. Formal experiments have a more obvious capacity for such an aim. However, it has been suggested that one of the disadvantages of formal experiments is that a finding of “no effect” cannot inform practitioners as to the reason why (and therefore the “researcher must go back to the drawing board” (Gorard *et al.* 2004: 586)). The current paper addresses whether and how, where a learning

theory *is* established (i.e. a set of falsifiable hypotheses), formal experimental educational research can both test this theory and inform practice, regardless of whether ‘an effect’ or ‘no effect’ is found.

This is addressed via a methodological discussion of a pair of experiments which aimed both to test a set of hypotheses about learning and to generate practical knowledge regarding a particular approach to the teaching and learning of foreign language grammar. Through discussion of these experiments, the paper suggests that certain contextual, theoretical and methodological conditions are necessary if an experiment is both to test a theory about learning and inform teaching practice. The paper acknowledges that achieving these conditions can be problematic and suggests ways in which the experiments under discussion may not have achieved them. Literature from the field of applied linguistics is drawn upon to provide examples of the kinds of issues and debates which require attention in such a study. Detailed substantive presentation of the illustrative study is beyond the scope of this paper (see Marsden 2004, 2005, 2006 for reports of the actual study for practitioner and researcher audiences). A *resumé* of the methods of the study is provided in section I. The rest of the paper discusses the study’s rationale and related ethical issues (section II), the theoretical framework and its implications (section III), the main findings and claims (section IV) and some lessons learnt (section V). Each section focuses on issues which relate to the study’s capacity to inform both educational practice and test a learning theory.

I *Resumé* of the methods

Design

Two experiments were carried out in two secondary schools in England to investigate two different ways of teaching French verb agreements and auxiliaries (henceforth,

verb inflections) to 13-14 year old (year 9) learners. Two teachers, via a series of professional contacts, agreed to participate and both schools selected top set classes to take part in the study. A questionnaire confirmed that participant learners did not have significant exposure to French outside the classroom, and had approximately 180 hours prior exposure to instruction.

In the experiments in both schools, one class (27 learners in school 1 and 29 in school 2) was divided into two equivalent groups using randomised matched pairs on the basis of their pretest results (Cohen, Manion & Morrison 2000: 216). Split class designs such as this are not common but can help to counter sampling bias due to the 'cluster effect' or 'inter-group correlation' (Torgerson & Torgerson 2003: 73, Moore *et al.* 2003: 679-80) i.e. the performance of children (or teachers) within the same class (or school) is more likely follow a similar pattern than that of children in a neighbouring class (or school). After the classes were split, one group received Processing Instruction (PI, as conceived by VanPatten, 1996, 2002 & 2004). The other group received Enriched Input (EI), an intervention designed specifically for this study, which reflected one element of common classroom practice and operationalised the 'alternative hypothesis' (discussed later). In the experiment in school 2, a parallel, intact class of 30 learners received only the tests (henceforth, the non-active control class). Allocation to PI and EI was 'blind' in that the learners were told they were being split into two smaller teaching groups of equivalent levels, which were called "blue" or "green". Attempts were made to reduce the effect of the 'teacher variable' by alternating the teaching of PI and EI between the regular teacher and the researcher.

The PI and EI interventions lasted for about eleven hours over about seven weeks. The timescale of the study is laid out in appendix 1. PI and EI are different types of

grammar pedagogy, consisting of reading and listening tasks only (no speaking or writing). PI is based on a set of hypotheses, called Input Processing, about what learners pay attention to and therefore, the theory suggests, learn, when they read or hear sentences in a foreign language.

Outcome measures

In both experiments, pretests, immediate post tests and delayed post tests (14-16 weeks after the intervention) were carried out to assess learners' progression in the accuracy of their interpretation and production of a range of verb inflections. Four measures were taken: listening and reading (interpretation), and writing and speaking (production). The battery of tests included a range of task types, including sentence completion, multiple choice, written and oral narratives and semi-structured conversations.

In addition to achievement test scores, attitudinal data was collected from the teachers and pupils, using questionnaires (pre and post intervention), interviews and focus groups. The close relationship between novelty and/or motivation and achievement has been demonstrated (for example, Burstall & *et al.*, 1974; Freedman 1978; Fitz-Gibbon & Reay, 1982 and for discussions of the Hawthorne effect see, e.g. Beard, Bligh & Harding, 1978; Borg & Gall, 1979; Cohen, Manion & Morrison, 2000; Brown, 2002). Attitudinal data was therefore collected to assess whether one intervention was preferred or considered more novel than the other. In addition to these steps, the act of creating an apparently 'new' comparison intervention (EI in the current study) was intended, amongst other things, to address the impact of novelty more satisfactorily than having 'normal teaching' as the only comparison.

Monitoring the process³

Lessons before, during and after the intervention (i.e. before the delayed post tests) were documented in each of the PI, EI and non-active control lessons, using mainly audio or video recordings, and, occasionally, field notes alone. One reason for documenting the teaching was to monitor the delivery of the interventions as they were happening in fairly natural (externally valid) environments. It is possible that monitoring itself may also have improved adherence to the protocol as teachers, although under no obligation, may be sensitive to the checking function of the observation. This observation data established fidelity of treatment between the pre and post tests in both schools i.e. PI and EI were carried out reliably, and teaching in the non-active control class did not focus on the target language forms. Between the post and delayed post tests in school 1 there was no further teaching of the target forms, so differences in scores at post and at delayed post test in school 1 could be ascribed to impacts of PI and EI. In school 2, differences in scores at post test in school 2 could also be ascribed to impacts of PI and EI. However, between the post and delayed post tests there was extensive teaching in school 2 of the target forms, thus affecting the capacity of the experiment in that school to ascribe causality to the experimental interventions alone at delayed post test.

Differences between the contexts of the two experiments

Process monitoring not only assessed the validity of the study, but was also required to assess whether certain interventions may have different impacts in particular contexts, and to help teachers determine for themselves the relevance of the study (see similar arguments for rich descriptions of case studies, e.g. Bassey, 1984; Hiebert *et al.*, 2002). Extensive monitoring of the two contexts revealed differences between the two contexts

that were relevant to the current study. Compared to learners in school 1, the learners in school 2 were more accustomed, for at least one year prior to and several months after the intervention, to tightly focused sequences of activities which taught and tested verb inflections, particularly for the purposes of accurate written production. At the outset, learners in school 2 were also at a statistically significantly higher developmental stage in terms of their competence in the area of French grammar under focus.

This *resumé* raises many contextual, ethical, theoretical and methodological issues relating to this study's capacity both to inform practice and test a learning theory. These issues are now discussed with reference to the illustrative study and the general educational and applied linguistics research literature.

II Contextual issues: Rationale and Ethics

i) Rationale

One obstacle which has been proposed as rendering educational experiments problematic is a lack of consensus regarding what education should strive to achieve and how. For example, Pring (2000) notes "different people will have different views of what is worth learning, and thus of what counts as a 'good education' or an 'educated person'. And that creates complications for educational research" (p. 14, see also p.22, p.158 and p.160; Hammersley 1997, 1999 & 2001; Wilson & Wilson, 1998; Atkinson, 2000). In order to inform practice and test a learning theory, there clearly needs to be some level of consensus regarding the nature of the 'problem' and 'solution' under investigation. In the current study, apart from the fact that studying a Modern Foreign Language (still usually French) is obligatory for year 9 learners,

further evidence for the relevance of the study, and the interventions, was sought from a range of stakeholders and commentators. These sources suggested that, despite some arguments to the contrary, the learning objective under focus (i.e. promoting accurate manipulation of verb inflections in French) is

a) regarded as a feature of progression which can emerge amongst learners at a similar stage to those in the current study (e.g. Myles, 2003; Rule & Marsden, in press),

b) perceived as problematic by a range of stakeholders (e.g. pupils (Lee, 1998; McPake, 1999), secondary teachers (Bryan, 2001; Hogg, 2001), higher education commentators (Klapper, 1997; Wright, 1999), GCSE and A level examiners (Hurman, 1992; Neil & Laverty, 2001), OFSTED inspectors (OfSTED, 2002; Dobson, 1998), policy makers (QCA, 2000; DfES 2003), and applied linguists (Chambers & Richards, 1995; Mitchell, 2000).

and c) that potential improvements to promoting the learning of grammar are worth pursuing (e.g. VanPatten, 1996; Harley, 1992; Doughty & Williams, 1998; Mitchell, 2000).

Such a battery of evidence from relevant groups (including non-interventionist studies) represents, it is argued here, sufficient consensus regarding the existence and nature of the educational problem and justifies exploring how to improve the situation.

Of course, the same problem of reaching a consensus about what learning is faces other effective teaching research designs. However, a caveat *is* required which is rarely articulated in effectiveness of intervention studies, i.e. that in another age, with another set of objectives for Modern Foreign Language education, improving ‘accuracy’ may be less relevant (see for example the scenarios presented in Mitchell, 2002). If claims

from studies are limited to a context where learning of x is deemed important, then education's sensitivity to socio-political-cultural climates does not render educational experiments unfeasible or inappropriate, and is not necessarily "bothersome" (Bassey 1984: 114), at least not in any practical sense. However, it *would* be bothersome for the purposes of testing theory if more open generalizations were not possible. This point is explored further in later sections.

ii) Ethical concerns

The quality and quantity of current provision in the relevant context is central to addressing ethical concerns about educational experiments, and therefore to assessing their capacity to inform practice and test theory. If current provision relevant to the experimental intervention (in this case, second language grammar pedagogy) is homogenous, it could be unethical to deny this standard provision to some learners (Hammersley, 2002: ch.1). However, in practice, despite the centralisation of education in the UK, the techniques, status and quantity of second language grammar pedagogy are largely context-(e.g. teacher-, school-, text book-) dependent (Mitchell & Hooper, 1992; Bygate, Tonkyn, & Williams, 1994). If experimental designs are refused on the grounds that differential treatment is unethical, then this creates the illogical position that differential teaching styles are acceptable if part of the *status quo*, but unethical if introduced systematically as part of an experiment.

Another obstacle which has been suggested as constraining educational experiments is that because there must be some evidence (including theoretical, pedagogical, or laboratory evidence) that an experimental intervention might be beneficial (in order to justify trialling it), it is unethical to deny the comparison learners of that potentially beneficial intervention (Kember, 2000; Gall, Borg, & Gall, 1996). However, in reality,

it is unlikely that there is unequivocal evidence of the positive (or negative) effects of any intervention. For example, see Torgerson & Torgerson (2001) for discussion of cases where intuitively appealing social interventions were actually found not to be beneficial once trialled, and systematic reviews suggest that it is almost impossible to endorse, unequivocally, a particular approach. In the case of foreign language grammar pedagogy, there are many sets of evidence suggesting potentially, but by no means unequivocally, effective interventions. For overviews of the potential impacts of a range of grammar pedagogy techniques see, for example, Norris & Ortega (2000), Mitchell (2000), Doughty & Williams (1998); for reviews supporting the specific intervention in the current study (PI), see VanPatten (1996, 2002, 2004); for more measured appraisals of PI, involving different comparison treatments and different outcome measures, see DeKeyser *et al.* (2002), Collentine (1998), Farley (2001a & b) and VanPatten & Sanz (1995). Furthermore, very little of this research has been carried out in the UK secondary school foreign language learning context, and so the impact (positive or negative) of these techniques on a range of relevant outcome measures cannot be known *before* an experimental research design is deemed appropriate.

III) Theoretical constraints: theory selection & its consequences

This section discusses the impact that the nature of the theory, and the use made of it, have on an educational experiment's capacity to both test that theory and inform practice.

Studies which use a 'theoretical framework' to develop teaching practice, but do not intend to test that theory, are fairly frequent in educational research. For example, Brown *et al.* (2003), Moore *et al.* (2003) and Black & Wiliam (2003) all base the

interventions used in their studies on theoretical underpinnings (on “Vygotskian theory”, “active learning” and “constructivist principles” respectively). However these studies were not designed to test a learning theory by setting up the necessary conditions to falsify it (Popper 1968).

Frequently, pedagogical interventions are multi-faceted and therefore their specific theoretical underpinnings are difficult to specify (see some of the RCTs reviewed in Torgerson & Zhu (2003) where, for example, the use of an ICT package in small groups removed from the main classroom could be compared to normal lessons, leaving ICT, small group work and/or removal from the classroom uncontrolled). Although “identifying distinct and standardised treatments in education” (Hammersley 1997, p.145) can be problematic, in fact, however, isolating ‘critical’ elements of teaching is not crucial if, in practice, a *whole* package would either be implemented or not. It is not contentious that experiments which evaluate either a-theoretical instructional packages or packages which could be explained by multiple hypotheses about learning can feasibly inform practice. However, to test a specific learning theory, the multi-faceted nature of pedagogical packages can be problematic.

This point can be illustrated by the interventions, and their theoretical underpinnings, used in the current study. PI includes a brief grammar explanation and two *different* types of input task. One of these task types (the ‘referential activities’) clearly operationalises the theory claiming to underpin it (Input Processing principles) and the other task type (a kind of Enriched Input) does not. However, previous PI studies have treated these two task types as ‘one intervention’, comparing it to a range of types of traditional, *output* based (speaking and writing) grammar teaching techniques (e.g.

VanPatten & Cadierno, 1993; Allen, 2000; Benati, 2001). In addition, these and other studies have not held the explicit grammar explanation component of PI constant across experimental and comparison groups. These complex comparisons meant that multiple variables could be the cause of differential learning gains. The nature of many of these comparisons was motivated by VanPatten & Cadierno's (1993) claim that 'traditional output practice' is the most common style of second language grammar pedagogy in the USA. Their research has therefore informed teaching practice rather than testing the specific hypotheses claimed to underlie PI. To attempt to achieve both of these, the current study compared two treatments in each of which only one element was different (i.e. the referential task type) and, as described below, this element operationalised the hypothesis being tested.

i) Brief exposition of theory (Input Processing) and pedagogical package (Processing Instruction).

Input Processing theory includes a set of hypotheses that suggest what type of language features learners normally attend to when hearing or reading a second language. The pedagogical technique is designed to *dissuade* learners from following these tendencies. It is beyond the purpose of this paper to present and critique this theoretical framework in any detail (see DeKeyser *et al.*, 2002; VanPatten, 2004; Marsden 2004 & 2006).

Relevant key hypotheses are as follows:

“Principle 1a: The Primacy of Content Words Principle. Learners process content words in the input before anything else.

Principle 1b: The Lexical Preference Principle. Learners will tend to rely on lexical items* as opposed to grammatical form to get meaning when both encode the same semantic information.”

VanPatten (2004: 14)

* *'lexical items' refers to nouns, pronouns, verb stems, adverbs and adjectives*

The hypotheses are falsifiable and significant in the terms suggested by Pring (2000): that is, acceptance of the hypotheses “makes a difference to how one views the world, understands experience or engages in practice” (p. 127). Accepting the truthfulness of these hypotheses entails accepting that learners, in the early stages of second language development, do not pay attention to verb inflections e.g. that in a sentence such as “*Hier j’ai joué au tennis*” [yesterday I played tennis], learners would use ‘*hier*’ to interpret ‘pastness’, rather than the verb inflection ‘*ai*’. Acceptance of the hypotheses also entails a belief that engaging in a particular kind of listening and reading activity (PI) would help learners to improve their interpretation and production of verb inflections. In the current study, these theoretical principles were operationalised for testing in the classroom as PI (reflecting the null hypothesis that the principles above, and their pedagogical implications, are valid) and Enriched Input (reflecting the alternative hypothesis, that the principles above are invalid). In summary, the comparison was:

Explicit grammar presentation + ‘PI referential activities’ + Enriched Input
compared to

Explicit grammar presentation + Enriched Input + Enriched Input

The same number and type of target verb inflections were in each of the PI and EI activities.

As shown above, of the three components in PI, components 1 and 3 were essentially held constant across both PI and EI. Component 1 is a brief (approximately 1 minute) explicit description of the grammar point. Component 3 (a kind of EI) consists of

listening and reading activities which contain an input flood of the target language forms (i.e. multiple exemplars) but which do not *force* learners to interpret the meaning of these forms, e.g.:

Tick if you think the following activities are normal, put ! if you think they are a bit weird:

1. Le bébé promène le chien

[The baby walks the dog]

2. Les professeurs corrigent les devoirs

[The teachers correct homework]

Other tasks within EI ask learners, for example, to order sentences, match them to pictures, express their preferences or opinions about them or state how relevant they are to their own lives.

Both components 1 and 3 are compatible with current practice (as can be seen by referring to classroom observation data e.g. Bygate *et al.*, 1994; Hogg, 2001; Marsden, 2005; current text books such as McNab, 1994). Adapting or simulating common teaching practice, obviously enhances a study's practical applicability⁴. It also reduces the likelihood of teachers abandoning the experimental protocol, if they consider the intervention irrelevant or ineffective (see, for example, Harley 1993).

The defining feature of EI listening and reading tasks is that learners can complete them without *having* to notice the verb ending. These activities therefore operationalise the alternative hypothesis (that the principles above are invalid), and would be evidenced by learning gains following EI. Learning gains would suggest that learners can pay

some level of attention to verb inflections whilst only actually having to pay attention to nouns, verb stems, adverbs and adjectives.

EI activities were compared to component 2 of PI (referential activities) such as:

According to the following sentences, underline who does the activity:

- | | |
|-------------------------------------|------------------------------|
| 1. <i>Le bébé / les parents</i> | <i>promène le chien</i> |
| [The baby / the parents | walks the dog] |
| 2. <i>Le chat / les professeurs</i> | <i>corrigent les devoirs</i> |
| [The cat / the teachers | correct the homework] |

The Input Processing principles above (claimed to underlie this teaching activity) predict that learners would choose, wrongly, that parents walk the dog. This is what learners would understand by only interpreting the lexical items in the sentence and relying on their world knowledge. PI claims that learners will learn language forms which are communicatively redundant (the ‘*ent*’ verb inflection above) more speedily if they are forced to interpret the *meaning* of such forms e.g. that ‘*ent*’ must refer to more than one person⁵.

In summary, the choice of theoretical framework had several important consequences for an experimental approach to be able to test a theory of learning and inform practice.

a) The theory could be operationalised as classroom activities, both in terms of the null and alternative hypotheses.

b) Both the experimental and comparison interventions adequately resembled some element of common teaching practice to render them sufficiently appealing to practitioners.

c) It was possible to make one activity type (the PI referential component) the comparative dimension, reducing the likelihood of multiple causality.

ii) *Apples and pears: the potential fallacy of comparing educational treatments*

Another constraint for educational experiments is that whatever an experimental treatment is compared to is highly likely to have (should have!) some educational benefits (unlike a placebo in a drug trial). Furthermore, given the limitations on curriculum time, changes in teaching practice entail one thing being sacrificed for another (rarely just supplemented, though see Moore *et al.* 2003). Under these circumstances, the definition of interventions which can be fairly compared becomes problematic, as two different activities are likely to have slightly different objectives, each entailing slightly different advantages and disadvantages. As argued above, in the current study there was some parity of aims between PI and EI in that both can be, and are, used to promote grammatical accuracy. Therefore using one set of measurements to evaluate the effectiveness of two different treatments was appropriate.

However, if the educational objectives of each treatment are not identical, then the effectiveness of the experimental treatment must also be assessed in terms of the aims (potential advantages) of the comparison treatment. PI and EI are used as an example. One of the participating teachers in the current study suggested in the post intervention interview that they could imagine using the EI style of activity as a *vocabulary* learning activity (also a good illustration of the necessity of multi-method approaches).

Therefore a *post-hoc* measure was taken to assess whether one intervention type was superior in terms of vocabulary learning by counting the number of lexical verbs produced in the writing and speaking post tests compared to the pre tests. Although the

results (see Marsden 2004) showed no significant benefits for either treatment type in terms of this measure, it is acknowledged that this was probably not a sufficiently rigorous measure, which should have assessed improvement in the range of verb types used and other grammatical categories such as nouns and adjectives. If the aim of experimental ‘what works’ studies is to develop our understanding of the different effects of different intervention types within finite curriculum time, then monitoring each intervention’s benefits and weaknesses is essential⁶.

This discussion has illustrated how, for an experimental study to test both a learning theory and inform practice, the experimental and comparison treatment shared some learning objectives and their potentially *different* impacts on learning could be assessed (though this assessment could have been improved in the current study).

IV) Discussion

In light of the above, and building on the *resumé* of the study in section II, the findings and claims from the current study are summarised, to illustrate the extent to which the study may and may not inform practice and test Input Processing principles. As noted earlier, a detailed presentation of the substantive findings from this study are beyond the scope of this paper and are, to a certain extent, irrelevant as the argument relates to the *potential* of a small scale study to test a learning theory and inform practice, regardless of the actual findings.

i) Could the study claim to inform practice (and does this have any impact on its capacity to test theory)?

Analysis of the findings from the study (see Marsden 2004, 2005, 2006) warranted the following claims.

Attitudinal data suggested that learners found both PI and EI activities new and enjoyable and all learners said they would like to continue to use them, though several with the caveat that use of them be less intensive⁷. No evidence was found to suggest that teachers or pupils preferred one intervention to another. For the sake of argument, it is noted that opposite findings would also have been useful, e.g. by providing one piece of evidence to dissuade teachers from adopting PI and/or EI. However, if teacher and/or pupil preference for one intervention had been found, this could have threatened the internal validity of the study for *theory* testing (requiring the isolation of variables) e.g. causal factors could then have involved motivation and/or the psycholinguistic learning principles.

The attitudinal data was also used to explore metacognition and participants' intentions. In the current study, it was found that there was a higher tendency for PI learners to explain why they liked PI with comments such as "I notice the small words more now" (see Marsden 2004). Such data can be important (crucial according to some, e.g. Pring, 2000:67-71) for tracking causal factors in educational research and can triangulate evidence for the same phenomenon (Gorard, 2002a & c; Frazer, 1995). However, full data collection and analysis of this kind requires larger and better resourced projects than the study reported here.

The achievement test results in school 1 showed that for learners who were in the very early stages of developing a verb inflection system, using input activities (PI) which

forced them to interpret the meaning of verb inflections resulted in statistically significantly larger learning gains than EI in all four measures used, including in delayed post tests up to 23 weeks after the intervention. EI learners did not make any gains in any of the measures (including once the normal teaching had resumed). This suggests that the learning objectives of activities which do not force learners' attention on the meaning of verb inflections (EI style) require clarification e.g. they may not be effective for learning verb inflections for learners at an early stage of developing verb inflections; other potential benefits of such activities (e.g. vocabulary learning) require further investigation.

In school 2, the listening and reading measures also indicated advantages for PI over EI at later developmental stages. The findings showed that for learners at more advanced stages of their emerging verb inflection system (and who were more accustomed to grammar-focussed sequences of activities), PI promoted the interpretation (reading and listening) of verb inflections, though to a lesser extent than for learners who were at an earlier developmental stage (and who were less accustomed to grammar-focussed sequences of activities). Methodological and contextual factors in this school rendered the speaking and writing measures uninterpretable⁸, and, as already noted, the delayed post test data.

One outcome of the study therefore, which was not intuitively appealing for the participant teachers at the outset, was that input practice *alone* resulted in gains not only in input based skills (listening and reading) but also in speaking and writing. For teachers who wish to try new methods, this study could provide ideas, and suggest their likely outcomes, in terms of attitude and achievement, depending on their learners' stage and experience of grammar teaching. This study, as it stands, may inform (not

dictate) practice (Beard, Bligh & Harding, 1978; Hammersley, 2004; Hodgkinson & Smith, 2004). It is emphasised that if different results had emerged from the study (e.g. that PI was consistently less than or equally effective as EI), then these too may have been able to ‘inform’ practice e.g. that listening and reading activities such as expressing opinions about content may also serve to promote grammatical accuracy regardless of developmental stage - a useful possibility in terms of efficient use of curriculum time.

On a larger scale, as there was no evidence in either school that PI had *negative* effects according to the measures taken, when compared to EI and to the non-active control, this small scale study could be used as one justification for a larger trial. The study’s findings, if replicated and combined with previous research in the area, could inform policy makers of the potential role of listening and reading in the development of learners’ productive grammars. Current statutory and non-statutory policy documents and pedagogical guides tend to promote listening and reading for other areas of language learning, such as gist comprehension, see DfEE (1999), DfES (2003), QCA (2000) and Turner (1995). Already, OfSTED (2004) report that teachers are making way for “the increased emphasis in the [KS3 (DfES 2003)] framework on teaching grammar” by “eschewing lengthy and ineffective listening comprehension exercises” (p.47), implying that there is a perceived tension between listening skills and developing grammar in MFL learning.

As mentioned earlier, attempts to reduce the effects of the teacher variable were made in this study by alternating the teaching between the regular teacher and the researcher. However, in a study aiming solely to inform practice, the teacher variable can be

desirable e.g. allowing teachers' interpretations of interventions to vary can facilitate some assessment of how variety in the delivery affects the impact of an intervention. However, as the current study aimed to test a learning theory, limiting the impact of the teacher variable was appropriate. This issue represents a considerable tension in studies both aiming to test a theory and inform practice.

There is another important constraint on the potential of this study to inform practice at a level greater than *ad hoc*, localised, adoption of the techniques. This constraint relates to the achievement outcome measures used. The debate surrounding the definition and measurement of knowledge (e.g. Pring 2000, and see Bachman & Cohen, 1998; Norris & Ortega, 2000; Truscott, 1998 & 2004 regarding the measurement of progression in second language proficiency) is not addressed here as it is relevant to any type of investigation of improvements in practice (e.g. Bassey 1984 regarding action research approaches). However, one related issue is specific to experiments which aim both to test a specific hypothesis and inform practice. For researchers wishing to inform *practice*, the use of standardised achievement tests as outcome measures can be appropriate so that findings may be compared to known population scores and their generalisability ascertained (Gorard, 2003). In contrast, in a study testing a specific learning hypothesis, standardized achievement tests are usually too blunt to measure relevant outcomes. In large scale research, this problem could be addressed by calculating appropriate sample sizes for the study (Torgerson & Torgerson, 2003). However, in a small scale study such as the one under discussion here, the lack of confidence about how well the samples reflect the population (e.g. all learners with 180 hours of French instruction) can limit the ease with which practitioners can gauge the relevance of the findings to their own context. In this

respect, the current study has three partially mitigating characteristics: i) some of the tests were adapted from those used in other grammar pedagogy and second language acquisition studies, and some of these studies could provide information about the production of verb inflections amongst other learners with approximately 180 hours of French instruction (e.g. Macrory & Stone, 1996 & 2000; Mitchell & Dickson, 1997; Hogg, 2001; Bryan, 2001; Myles, 2003; Rule & Marsden, in press); ii) the four measures are very broadly reflective of current indicator types used in the National Curriculum and GCSE (listening, reading, speaking and writing); and iii) the actual tests and learners' scores can be made available to practitioners if they wish to check the relevance of the study's samples to their own situation.

ii) Can the study claim to inform theory?

Although the actual findings presented here generally support the Input Processing hypotheses presented earlier, the reader is invited to imagine alternative scenarios which would have countered their validity.

In school 1, support was found for VanPatten's theory of input processing according to both the interpretation and production measures.

In school 2, support was found for the theory according to the interpretation measures⁹. However, the effect sizes of PI compared to EI were smaller than those found in school 1 (and to those in a relevant meta-analysis (Norris & Ortega, 2000)). The learning gains made by the EI learners could indicate that learners at higher developmental stages *are* more able than learners at lower developmental stages to learn verb inflections even when the verb inflections are communicatively redundant. These

results may inform one of the sub principles in Input Processing theory, i.e. how higher developmental stages may affect learning processes. VanPatten (2004) suggests, as a caveat to the principles 1a) and b) above:

“Principle 1e: The Availability of Resources Principle. For learners to process either redundant meaningful grammatical forms or nonmeaningful forms, the processing of overall sentential meaning must not drain available processing resources.”

This sub principle would suggest that EI activities may promote learning of verb inflections if learners have sufficient ‘processing resources’. In order to define the term ‘processing resources’, three characteristics of learners in school 2 suggest some areas research could explore: i) more efficient use of explicit grammatical rules when interpreting input; ii) more established initial representation of verb inflections; and iii) more efficient interpretation of vocabulary. One or all of these variables could facilitate the learning of verb inflections during tasks such as EI. As a result of these findings, the current study moved into a phase of laboratory research to assess the existence and nature of implicit learning of verb inflections amongst learners of different developmental stages.

V) Conclusion: further research and development

i) Summary

Given the special regard the experiment appears to have for informing practice and testing theory, this paper evaluated the methodology of an experiment which aimed to do both, and examined the nature of the knowledge it generated. It was argued that to justify offering (necessarily) different interventions, certain contextual conditions were

required e.g. 'reasonable consensus' amongst stakeholders regarding the broad nature of the educational problem (relating to both the theory and practice), and non-homogenous current provision in the area of teaching under focus. It was also argued that experimental classroom studies which aim to test a specific learning theory *and* inform practice raise additional challenges to those posed by educational experiments evaluating packaged or holistic interventions to inform practice alone. For example, it must be possible to operationalise the learning theory in question as classroom practice, both in terms of null and alternative hypotheses, and these classroom practices must bear some resemblance to current practice. For the study to inform practice where there is finite curriculum time and externally imposed learning objectives, both the potential benefits and disadvantages of each intervention need to be visible in the outcome measures (assuming that the interventions being compared do not have identical learning objectives). It was suggested that randomised split class designs and monitoring the context (including impacts of the experimental conditions on attitudes and motivation) were useful for exploring causal relationships.

ii) Lessons from the study

a) It has been acknowledged that the outcome measures used, although appropriate for testing the theory, could be more helpful if the relevant population scores were known, for practitioners and policy makers wishing to ascertain the generalisability of the findings. Besides replicating small scale studies, an additional solution to this problem is to continue to build and use large databases documenting learner progression (e.g. Myles & Mitchell (2004) for French, and Brown *et al.* (2003) for maths). Another solution is to include in the outcome measures some standardised tests, alongside the measures needed for testing the specific learning theory, though this may increase demands on curriculum time devoted to the research.

b) It has been frequently suggested that one way of increasing the relevance of educational research to practice is to increase the involvement of teachers in it (e.g. Hiebert *et al.*, 2002 and various chapters in Thomas & Pring, 2004). It has also been noted, however, that simply increasing practitioner or political control over academic research is unlikely to impact on the *quality* of the research (e.g. Gorard, 2002b). Notwithstanding this note of caution, the conduct of the current study has suggested specific ways in which more involvement from teachers could lead, in the case of experimental studies, to improvements in research quality, strengthening the potential of the study to inform both theoretical and practical issues. Three examples are offered here.

First, as teachers often make decisions about ‘how’ to teach something (given that ‘whether or not’ to teach it is, currently, often out of their control), teachers’ opinions about the objectives of the different interventions can help to ensure that fair and valid comparisons are made and that outcome measurements are selected to reflect those objectives. In the current study, negotiation about this before the pretests would have led to an additional outcome measurement being taken.

Second, in the current study, reliable delivery of the interventions may have been a consequence of the strong rationale for the current study (i.e. an area of education, foreign language grammar pedagogy, well-known for being problematic). Other reasons could have been at play including: a desire to help the local university; a ‘side effect’ of monitoring the teaching; the size of the split class teaching groups may have encouraged reliable participation, both initially and subsequently; the fact that the

teachers were confident that their 'top ability' learners would not suffer from involvement in the project. It is suggested that teachers' reasons for their involvement in such a project require in-depth exploration before the study, as the study's internal validity will rely on the strength and nature of those reasons.

Third, and related to the above, is that increasing teachers' ownership of the project (e.g. increasing their involvement in the design of the materials and protocol) may also help adherence to the protocol. Three illustrations of this in the current study are given here. i) The speaking pretest measure in school 2 was probably invalid due to the related teaching which was going on in the classroom during the pretests (increased teacher ownership may have pre-empted this parallel teaching and testing). ii) One of the participating teachers' concerns was that although the topics (e.g. 'hobbies', 'daily routines') through which the verb inflections were taught, were in line with the teachers' own schemes of work, the order of the topics had to be different (teacher involvement in the materials design may have reduced this problem). iii) In one school, measurement of longer term impacts of the interventions was not possible because teaching relevant to the test was carried out before the delayed post test, therefore the experimental intervention may not have been the cause of any differences found at delayed post test (this is not necessarily problematic for studies informing only practice, but it is problematic for studies which claim to be theoretically-informed).

These points clearly illustrate the classic "experimenter's dilemma" Jung (1971) i.e. the trade-off between maintaining control over the protocol (i.e. maintaining internal validity, perhaps particularly necessary in a study which aims to test a theory) and reducing the artificiality of an experiment (i.e. maintaining external validity, perhaps

particularly necessary in an experiment which aims to inform practice). This tension is arguably greater in an experiment which is aiming to test a learning theory *and* inform practice. Given that currently (rightly or wrongly) researchers are more likely to be familiar with the related research agenda than teachers, some control by them over the nature of the intervention would remain necessary in such studies. Combinations of academic research with teacher education and/or professional development could be one avenue for resolving some of these tensions. However, systems to give teachers time and incentive to collaborate with researchers in such projects are, at present, limited in their capacity to co-ordinate or fund activity which moves *beyond* simply “increasing the proportion of educational research that is carried out by practising teachers” (Hammersley 1997: 150). Efforts are required to avoid mistakes from the past (e.g. see Webb, 1990 for a discussion of the perceived transmission of supposed good practice from researchers to practitioners). However, in an environment where genuine enquiry and professional development opportunities are foregrounded, these mistakes do not mean that collaborative experimental approaches could not usefully be re-habilitated.

iii) Limits of this paper

This paper has not discussed in depth several of the problems which have come to be frequently associated with educational experiments, as they are in fact relevant to any effectiveness-of-intervention study e.g. the epistemological perspectives required to enable a researcher to draw on different data types, to delineate educational interventions and to make measurements of learning.

The generalisability of the arguments presented in this paper is probably limited. One reason for this could be that the learning objective of the illustrative study (i.e. the development of an accurate verb system) is unique as an educational objective. For example, within other areas of second language competence (e.g. sociocultural knowledge) or other educational enterprises (e.g. music, mathematics, citizenship education), debate about the nature of learning objectives and outcome measures may pose more of a problem for the feasibility and usefulness of experimental designs.

Another limitation to the generalisability of the arguments presented here is that there are probably only certain types of theory that can be tested in classroom experiments that also aim to inform teaching practice. Theoretical debates can be counter-intuitive and operationalising them in terms of classroom activities or outcome measurements may not be feasible or useful to practitioners⁹.

Another potential challenge to the arguments presented in this paper is whether the size of the contribution such studies can make to theory building and practice merits the investment required, but this perhaps simply echoes calls for caution regarding what educational research can promise (Pring, 2000; Hammersley, 2001 & 2003). It is noted, however, that some of the arguments in this paper are bound by the current socio-political climate. In another era (e.g. where teachers were enabled to adopt research-oriented approaches, open to public scrutiny), educational experimentation which can formally test both teaching and learning theories may not be constrained by some of the issues raised here.

References

- Allen, L. (2000). Form-meaning connections and the French causative. *Studies in Second Language Acquisition*, 22, 69-84.
- Andrews R, Torgerson C, Beverton S, Freeman A, Locke T, Low G, Robinson A, Zhu D (2004) The effect of grammar teaching (sentence combining) in English on 5 to 16 year olds' accuracy and quality in written composition. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.
- Atkinson, E. (2000). Critical dissonance and critical schizophrenia: the struggle between policy delivery and policy critique, *Research Intelligence*, vol. 73, 14-17.
- Bachman, L. & Cohen, A. (1998). *Interfaces between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Papers.
- Bassey, M. (1984) "Pedagogic research: on the relative merits of search for generalization and study of single events". In J. Bell, Bush, T. Fox, A. Goodey, J. and Goulding S. (Eds) *Conducting small-scale investigations in educational management* London: Harper & Row
- Bassey, M. (1995) *Creating Education through Research: a global perspective of educational research for the 21st Century* Newark: Kirklington Moor Press, in association with the British Educational Research Association.
- Beard, R, Bligh, D & Harding A. (1978) *Research into Teaching Methods in Higher Education* 4th edition. Surrey: Society for Research into Higher Education
- Benati, A. (2001). A comparative study of the effects of processing instruction and output-based instruction on the acquisition of the Italian future tense. *Language Teaching Research*, 5, 95-127.
- Black, P. & Wiliam, D. (2003). 'In praise of educational research': formative assessment, *British Educational Research Journal*, 29, 623-638.
- Brown, A. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings *The Journal of the Learning Sciences*, 2, 141-178.
- Brown, M., Askew, M., Rhodes, V., Denvir, H. & Wiliam, D. (2003). Characterising individual and cohort progression in learning numeracy: results for the Leverhulme 5-year longitudinal study. Paper presented at the American Education Research Association Conference. Chicago, April 21-25, 2003.
- <http://www.kcl.ac.uk/depsta/education/papers/AERA2003.pdf>, accessed May 2005.

- Bruner, J. (1990). *Acts of Meaning*. Cambridge, Mass.: Harvard University Press.
- Bryan, V. (2001). *Effective MFLs Grammar Pedagogy: Teaching and learning the present and perfect tense* Best Practice Research Scholarships Level 2 Report, <http://www.teachernet.gov.uk/professionaldevelopment/resourcesandresearch/bprs/>, accessed May 2005.
- Burstall, C. & et al. (1974). *Primary French in the Balance*. Windsor: NFER.
- Bygate, M., Tonkyn, A., & Williams, E. e. (1994). *Grammar and the Language Teacher*. Prentice Hall.
- Chambers, F. & Richards, B. (1995). The 'free conversation' and the assessment of oral proficiency. *Language Learning Journal*, 11, 6-10.
- Chambers, J. (1992). *Empiricist research on teaching: a philosophical and practical critique of its scientific pretensions*. London: Kluwer Academic.
- Cohen, L., Manion, L., & Morrison, K. (2000). *Research Methods in Education*. (5th ed.) London: Routledge Falmer.
- Collentine, J. (1998). Processing instruction and the subjunctive. *Hispania*, 81, 576-587.
- Cook, T. D. & Payne, M. R. (2002) Objecting to the objections to using random assignment in educational research. In F. Mosteller & R. Boruch (Eds) *Evidence Matters: randomized trials in education research*, (pp. 150-178). Washington, DC, Brookings Institution Press.
- DeKeyser, R., Salaberry, M., Robinson, P., & Harrington, M. (2002). What gets processed in processing instruction: a commentary on Bill VanPatten's "Processing Instruction: Update". *Language Learning*, 52, 805-823.
- DfEE (1999), *Modern Foreign Languages: The National Curriculum* DfEE.
- DfES (2003). *Framework for teaching modern foreign languages: Years 7, 8 and 9*. DfES.
- Dobson, A. (1998). *MFL Inspected: Reflections on inspection findings 1996-7*. CILT.
- Doughty, C. & Williams, J. (1998). *Focus on Form in Classroom Second Language Acquisition*. Cambridge University Press.
- Evans, J. & Benefield, P. (2001) Systematic reviews of educational research: does the medical model fit? *British Educational Research Journal*, 27, 527-541.
- Farley, A. (2001a). Authentic processing instruction and the Spanish subjunctive. *Hispania*, 84, 289-299.

- Farley, A. (2001b). The effects of processing instruction and meaning-based output instruction. *Spanish Applied Linguistics*, 5, 57-94.
- Fitz-Gibbon, C. & Reay, D. (1982). *Peer tutoring: brightening up FL teaching in an urban comprehensive school*. Publication no. 46. Curriculum, Evaluation and Management Centre, University of Durham.
- Foster, P. (1999) "Never mind the quality, feel the impact": a methodological assessment of teacher research sponsored by the Teacher Training Agency *British Journal of Educational Studies*, 47, 380-398.
- Frazer, E. (1995) What's new in the philosophy of science? *Oxford Review of Education* 21, 267-281.
- Freedman, E. (1978). *Experimentation into foreign language teaching methodology*. Unpublished PhD thesis, University of Essex.
- Gall, M., Borg, W., & Gall, J. (1996). *Educational Research: An Introduction*. Longman.
- Gorard, S. (2002a) Can we overcome the methodological schism? Four models for combining qualitative and quantitative evidence *Research Papers in Education* 17, 345-361.
- Gorard, S. (2002b). Political Control: A Way Forward for Educational Research? *The British Journal of Educational Studies*, 50, 378-389.
- Gorard, S. (2002c) The role of causal models in education as a social science? *Evaluation and Research in Education* 16, 51-65.
- Gorard, S. (2003) *Quantitative methods in social science: the role of numbers made easy*, London: Continuum.
- Gorard, S. with Taylor, C. (2004) *Combining Methods in Educational and Social Research* Maidenhead: Open University Press.
- Gorard, S., Roberts, K., & Taylor, C. (2004). What kind of creature is a design experiment? *British Educational Research Journal*, 30, 577-590.
- Hammersley, M. (1997). Educational Research and Teaching: a response to David Hargreaves' TTA lecture. *British Educational Research Journal*, 23, 141-161.
- Hammersley, M. (1999). Some Reflections on the Current State of Qualitative Research. *Research Intelligence*, 70, 16-18.
- Hammersley, M. (2001) On systematic reviews of research literatures: a 'narrative' response to Evans and Benefield, *British Educational Research Journal*, 27, 543-54.

- Hammersley, M. (2002) *Educational research, policymaking and practice*, London: Paul Chapman.
- Hammersley, M. (2003) Can and should educational research be educative? *Oxford Review of Education*, 29, 3-25.
- Hammersley, M. (2004). Some questions about evidence-based practice in education. In G. Thomas & R. Pring (Eds.), *Evidence-Based Practice in Education* (pp. 133-149). Maidenhead: Open University Press.
- Harley, B. (1992). Patterns of second language development in French immersion. *French Language Studies*, 2, 159-183.
- Harley, B. (1993). Instructional strategies and SLA in early French immersion. *Studies in Second Language Acquisition*, 15, 245-259.
- Hiebert, J., Gallimore, R. and Stigler, J. (2002) A knowledge base for the teaching profession: What would it look like and how can we get one? *Educational Researcher* 31, 3–15.
- Hodkinson, P. & Smith, J. (2004). The relationship between research, policy and practice. In G. Thomas & R. Pring (Eds.), *Evidence-Based Practice in Education* (pp. 150-163). Maidenhead: Open University Press.
- Hogg, I. (2001). *Effective MFLs Grammar Pedagogy: the teaching and learning of German verbs* Best Practice Scholarships Level 2 Report, <http://www.teachernet.gov.uk/professionaldevelopment/resourcesandresearch/bprs/> accessed May 2005.
- Hurman, J. (1992). Performance in the A level speaking test by candidates with GCSE training: oral examiners' views. *Language Learning Journal*, 5, 8-10.
- Jung, J. (1971). *The experimenter's dilemma*. London: Harper and Row.
- Kelly, E. and Lesh, R. (2002). Understanding and explicating the design experiment methodology, *Building Research Capacity*, 3, 1-3
- Kember, D. (2000), *Action Learning and Action Research: Improving the Quality of Teaching and Learning* Kogan Page.
- Klapper, J. (1997). Language learning at school and university: the great grammar debate continues (1). *Language Learning Journal*, 16, 22-27.
- Lather, P. (2004) Scientific research in education: a critical perspective, *British Educational Research Journal*, 30, 759-772.
- Lee, J. (1998). *The invisible child*. London: CILT.

- Macrory, G. & Stone, V. (1996). Knowing and using the perfect tense in French. In G. Blue & R. Mitchell (Eds.), *Language and Education: Papers from the Annual Meeting of the British Association for Applied Linguistics* (pp. 107-118). Multilingual Matters.
- Macrory, G. & Stone, V. (2000). Pupil progress in the acquisition of the perfect tense in French: the relationship between knowledge and use. *Language Teaching Research*, 4, 44-82.
- Marsden (2004). *Teaching and Learning of French Verb Inflections: A Classroom Experiment Using Processing Instruction*, Unpublished PhD thesis, University of Southampton.
- Marsden, E. (2005). Input-based grammar pedagogy: a comparison of two possibilities *Language Learning Journal*, 31, pp 9-20.
- Marsden, E. (2006). "Exploring Input Processing in the classroom: An experimental comparison of Processing Instruction and Enriched Input", under review *Language Learning*.
- McNab, R. (1994). *Avantage 3* Heinemann Educational.
- McPake, J. (1999). *Foreign Language in the Upper Secondary School: A study of the causes of decline* Scottish Council for Educational Research, Report No. 91.
- Mitchell, R. (2000). Applied linguistics and evidence-based classroom practice: the case of foreign language grammar pedagogy. *Annual Review of Applied Linguistics*, 281-303.
- Mitchell, R. (2002). Foreign Language Education in an Age of Global English. *Centre for Language in Education Occasional Paper No. 57* University of Southampton.
- Mitchell, R. & Dickson, P. (1997). Progression in Foreign Language Learning. *Centre for Language in Education Occasional Paper No.45* University of Southampton.
- Mitchell, R. & Hooper, J. (1992). Teachers' views of language knowledge. In C. James & P. Garrett (Eds.), *Language Awareness in the Classroom* (pp. 40-50). Longman.
- Moore, L. (2002). 'Lessons from using randomised trials in health promotion', *Building Research Capacity*, 1, 4-5.
- Moore, L., Graham, A. & Diamond, I. (2003) On the feasibility of conducting randomised trials in education: a case study of a sex education intervention, *British Educational Research Journal*, 29, 673-689.
- Morrison, K. (2001) Randomised controlled trials for evidence-based education: some problems in judging 'what works', *Evaluation and Research in Education*, 15, 69-83.

- Myles, F. (2003). The emergence of morpho-syntactic structure in French L2. In J. Dewaele (Ed.), *Focus on French as a Foreign Language Multilingual Matters*.
- Myles, F. & Mitchell, R. (2004). *French learner language oral corpora*
<http://www.flloc.soton.ac.uk/>
- Neil, P. & Lavery, R. (2001). An analysis of pupils' writing skills in the French GCSE examination. *Francophonie*, 24, 12-17.
- Norris, J. & Ortega, L. (2000). Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Language Learning*, 50, 417-528.
- Oakley, A. (2000) *Experiments in Knowing: gender and method in the social sciences* Cambridge, Polity Press.
- Oakley, A. (2001) Making evidence-based practice educational: a rejoinder to John Elliott, *British Educational Research Journal*, 27, 575-6.
- Oakley, A. (2002) Social science and evidence-based everything: the case of education *Educational Review*, 54, 277-286.
- OfSTED (2002). *Ofsted Subject Reports: Secondary Modern Foreign Languages 2000-2001* OFSTED.
- OfSTED (2004), *The Key Stage 3 Strategy: evaluation of the third year* OfSTED.
- Parry-Langdon, N., Bloor, M., Audrey, S. & Holliday, J. (2003) Process evaluation of health promotion interventions, *Policy and Politics*, 31, 207-216.
- Pawson, R. & Tilley, N. (1997) *Realistic Evaluation* London: Sage.
- Popper, K. 1968, *The Logic of Scientific Discovery* (2nd ed), London: Hutchinson.
- Prévost, P. & White, L. (2000). Missing surface inflection or impairment in second language acquisition? Evidence from tense and agreement. *Second Language Research*, 16, 103-133.
- Pring, R. (2000). *Philosophy of educational research*. London: Continuum.
- QCA (2000). *Schemes of Work for Key Stage 3 French* QCA.
- REEL (2004). *A systematic review of the characteristics of effective foreign language teaching to pupils between the ages 7 and 11* http://eppi.ioe.ac.uk/EPPIWebContent/reel/review_groups/TTA/MFL/MFL_2004summary.pdf, accessed January 2005.
- Rule, S. & Marsden, E. (in press). The status of functional categories in the acquisition of negation in early French L2 grammars. *Second Language Research*.
- Shadish, W., Cook, T. & Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

- Shavelson, R., Phillips, D., Towne, L. & Feuer, M. (2003). On the science of education design studies, *Educational Researcher*, 31, 29-31.
- Slavin, R. (2002). Evidence-based educational policies: transforming educational practice and research, *Educational Researcher*, 31, 15-21
- Stenhouse, L. (1980) The study of samples and the study of cases, *British Educational Research Journal*, 6, 1-6.
- Stoddard E. & Renzulli J. (1983) Improving the writing skills of talent pool students. *Gifted Child Quarterly* 27, 21–27.
- Thomas, G. (2002) Theory's spell – on qualitative inquiry and educational research *British Educational Research Journal*, 28, 419-434.
- Thomas, G. (2005) *Strangers in paradigms: education and theory in the twentieth century*. Maidenhead: Open University Press.
- Torgerson, C. & Torgerson, D. (2001). The need for randomised controlled trials in educational research, *British Journal of Educational Studies*, 49, 316–329.
- Torgerson, C. J. & Torgerson, D. J. (2003). The design and conduct of randomised controlled trials in education: lessons from health care *Oxford Review of Education*, 29, 67-80.
- Torgerson, C. & Zhu, D. (2003). A systematic review and meta-analysis of the effectiveness of ICT on literacy learning in English, 5-16. *REEL, EPPI review centre*, <http://eppi.ioe.ac.uk/>. accessed May 2005.
- Truscott, J. (1998). Noticing in second language acquisition: a critical review. *Second Language Research*, 14, 103-135.
- Truscott (2004) The effectiveness of grammar instruction: Analysis of a meta-analysis, *English Teaching and Learning* 28, 17-29.
- Turner, K. (1995), *Listening in a Foreign Language: a skill we take for granted?* Pathfinder 26, CILT.
- Tymms, P. & Fitz-Gibbon, C. (2002) Theories, hypotheses, hunches and ignorance, *Building Research Capacity*, 2, 10–11.
- VanPatten, B. (1996). *Input Processing and Grammar Instruction in Second Language Acquisition*. Norwood, N.J.: Ablex.
- VanPatten, B. (2002). Processing Instruction: An Update. *Language Learning*, 52, 755-803.
- VanPatten, B. ed. (2004). *Processing Instruction: Theory, Research, and Commentary*. Mahwah, New Jersey: Lawrence Erlbaum.

- VanPatten, B. & Cadierno, T. (1993). Input Processing and Second Language Acquisition: A role for instruction. *The Modern Language Journal*, 77, 45-57.
- VanPatten, B. & Sanz, C. (1995). From input to output: processing instruction and communicative tasks. In F.Eckman, D. Highland, P. Lee, J. Mileham, & R. Rutowski Weber (Eds.), *Second Language Acquisition Theory and Pedagogy* (pp. 169-185). Lawrence Erlbaum Associates.
- Webb, R. (Ed.) (1990) *Practitioner Research in the Primary School*. Basingstoke: The Falmer Press
- Wilson, J. & Wilson, N. (1998) The subject-matter of educational research, *British Educational Research Journal*, 24, 355–363.
- Wright, M. (1999). Grammar in the languages classroom: findings from research. *Language Learning Journal*, 19, 39.

Appendix 1. Timescale of the study in weeks

	<i>Pretests</i>	<i>Intervention¹⁰: PI and EnI</i>	<i>Post tests</i>	<i>Normal instruction</i>	<i>Dp tests</i>
<i>Duration</i>	2	7	2	12	2
<i>Cumulative timescale</i>	2	9	11	23	25

¹ The pair of studies reported and the dissemination of their findings were funded by grants from the Economic and Social Research Council of Great Britain (R42200124319 and PTA-026-27-0252). Many thanks go to the teachers and pupils who participated in the research, and to **Rosamond Mitchell, Christopher Brumfit and Stephen Gorard for useful discussion about some of the issues raised**. Details about the actual study and analysis of the substantive findings can be found elsewhere (Marsden 2004, 2005, 2006).

² It is emphasized that this article does not discuss whether practitioner research could, given the right environment, contribute to the development of educational theory (e.g. Pring, 2000 ch. 6; Hiebert *et al.* 2002).

³ For discussion about combining qualitative and quantitative data see Moore (2002), Pawson & Tilley (1997), Parry-Langdon *et al.* (2003); Gorard (2002a) (though only 4 per cent of studies actually do so, according to an analysis by Gorard with Taylor, 2004: 141).

⁴ After the intervention, both participating teachers agreed that EI was familiar in that it contained exemplars of a target feature but did not focus learners' attention on that feature.

⁵ This does not involve the same processes as other grammar teaching activities where learners simply note that a certain verb inflection has been used or produce a verb inflection to agree with the subject.

⁶ For example, Moore *et al.* (2003) did not assess the potential disadvantages of their intervention, as their secondary measures did not take explore whether education about emergency contraception may increase the amount of unprotected intercourse. In contrast, Stoddard & Renzuli (1983) do attempt to take fit-for-purpose measurements, and their claims of effectiveness therefore can actually take into account different educational goals i.e. syntactic maturity versus creativity. However, it is noted that this issue could have consequences for another aspect of experimental designs i.e. measures assessing the advantages of both experimental and comparison groups may not necessarily correlate positively and so using these measures to create equivalent groups at the outset of an experiment might be inappropriate.

⁷ The questionnaire was sensitive enough to detect a decline in preference for French in the non-active control class who had a supply teacher after the post tests.

⁸ Extraneous variables could account for the production measures: the speaking pretests were probably invalid as a baseline score because the PI and EI learners were withdrawn from a French lesson in which the target features of the test were being taught, thus probably producing artificially high pretest scores making differential learning gains difficult to detect at post test (learning gains were similar across PI, EI and non-active control learners). The writing scores could probably be attributed to the sensitivity of the writing measure to the explicit grammar instruction component in PI and EI given school 2 learners' familiarity with written grammar teaching and testing.

⁹ To illustrate by returning to the area of language learning, some linguists consider that the absence of a verb inflection does not necessarily indicate that learners haven't gained some representation of the language feature (e.g. Prévost & White, 2000; Rule & Marsden, in press). However, incorporating such theoretical possibilities in outcome measurements would reduce a study's obvious relevance to practice, as it would entail stakeholders accepting '*il ne jouer pas au football*' as grammatically accurate (rather than '*il ne joue pas*').

¹⁰ Even if pupils missed some of the intervention, all pre, post and delayed post test measurements that were collected were included in the final analysis (see Torgerson & Torgerson (2003) regarding the need to monitor the impact of 'intention to teach/treat' in social interventions).