phase-6 AG

Analyses and Results of Initial Research Studies

An Independent Report

7 January 2009

Catherine Fritz
Lecturer
Department of Educational Research
Lancaster University
Lancaster, LA1 4YL

Don Passey
Senior Research Fellow
Department of Educational Research
Lancaster University
Lancaster, LA1 4YL

Peter Morris
Professor of Psychology
Department of Psychology
Lancaster University
Lancaster, LA1 4YL

CONTENTS

1. EXECUTIVE SUMMARY

1.1 Educational concerns and uses of technology

- Many technology facilities are available to support learners. So, why would learners, teachers or educators consider using yet another technology facility?
- Teachers and educators are concerned with supporting wide and deep levels of understanding, but recognise the need for understanding to be based on appropriate levels of memorisation of facts, ideas or events, rather than just on processes concerned with the more fluid aspects of application.
- The software facility considered in this research study, phase-6, focuses on supporting memorisation, through retrieval and recall practice. This technology is worthy of consideration, therefore, by teachers or educators concerned with supporting long-term memorisation.

1.2 Early technology-based resources and the development of research approaches

- Technology-based resources introduced into schools some 20 years ago were often quite small, specifically focused resources that supported specific aspects of learning or subject knowledge. The adoption of these resources was sometimes accompanied by research studies to identify impacts on learning. The methods often compared test outcomes arising from technological intervention to those where there was no technological intervention (using parallel control groups, easily possible with the generally low levels of technology available).
- A number of these controlled studies were set up, more often in the United States (US) than in the United Kingdom (UK). Over time, larger technology-based resources were developed, and their impacts were explored through research studies. Some of these large US resources, such as large-scale integrated learning systems (SuccessMaker, Jostens or Plato), for example, were introduced into the UK, and independent studies in the UK were conducted to look at impacts.
- Since those earlier studies were conducted (from about 10 years ago), there has been an enormous increase in ranges and levels of technology accessible to learners, both inside and outside schools. Identification of specific outcomes from specific technologies has, as a consequence, been made increasingly difficult.
- Recently, investigation of impact has often been undertaken and reported at a more general level, providing more widely conceived indicators of outcome, rather than specific indicators focusing on impact upon particular aspects of learning.
- Some wide research and evaluation studies have reported impacts on learning arising from uses of information and communication technology (ICT). Using subject attainment tests as measures of impact, some age ranges and subject areas have been found to benefit from technology involvement (at levels of statistical significance). The roles of school and classroom management, teacher involvement and pedagogy have been identified in parallel as important factors.
- When exploring impacts of specific technologies (rather than impacts arising from the entire range of technologies accessible to learners), a number of recent studies have focused on and investigated qualitative impacts. However, some recent UK-based quantitative studies have explored impacts of a specific technology on subject attainment (including a number undertaken by the Fischer Family Trust on SAM Learning online revision materials, resources which are often used outside as well as inside school).

1.3 Concepts behind phase-6 development and recent associated retrieval practice research

- Studies into forms of memorisation, retrieval and practice approaches undertaken and reported by Ebbinghaus, and the development of a box system to aid retrieval and practice approaches by Leitner, have provided the development background for the software facilities offered by phase-6. Prior to the development of this software, pupils in mainland Europe used boxes, divided into five spaces, to support revision and memorisation of words and phrases. phase-6 adapts this concept, changing the space or distance perspective into a time perspective.

- Considerable recent research in the US and the UK has looked at the benefits that can arise from appropriate uses of spacing retrieval practice. Expanding retrieval practice has been shown to be effective in many situations, not only for students, but also for normal adults, pre-school and older children, and the elderly.
- Although appropriate spacing is always beneficial for long-term learning, the use of increasing space intervals to support more effective retrieval and practice has been shown to be successful in some, but not necessarily in all, cases explored. One factor that affects learning and interacts with the type of spacing is access to feedback. So, when re-presentation of correct information is provided as feedback, for example, fixed spacing is as beneficial as expanding spacing.

1.4 Preliminary studies to explore impacts of phase-6 on learning

- Two preliminary studies looking at uses and impacts of phase-6 in schools have been undertaken. There were two maim aims for these studies. The first aim was to explore how investigations might be approached, in order to effectively identify impacts on learning, to draw conclusions that would offer recommendations for robust future studies. The second aim was to explore whether data gathered in two different contexts would identify impacts on memorisation when using the software facilities.
- The first aim of the research has been met in full. These studies have allowed the research team to identify a future approach for the investigation of impact. Recommendations concerned with this approach are offered at the end of this summary and in the conclusion at the end of the report.
- The second aim of the research was also met, but the picture of impact that was provided by the data from the two schools was not always crisp. However, it should be noted that across the entire range of studies, there was no indication that uses of phase-6 were in any way leading to negative impact; on the contrary, impacts identified were positive or neutral, and in some cases impacts with statistical significance were identified. The reasons for lack of clarity, and the impacts where statistical significance were identified, are highlighted in this summary and throughout the report.

1.5 A study in a California school

- 172 students in a school in California used phase-6 to support their learning of Spanish language vocabulary. They used phase-6 for the practice of some words, but not others. The study was undertaken by phase-6 in California, in conjunction with a selected school, and the test results were given to the research team.
- An analysis of test results indicated positive impact, both at the level of improved performance, and of improved prediction of performance. However, it was not possible to draw a firm conclusion that the differences in performance identified were due to phase-6 alone. Further data about the comparative difficulty levels of the two sets of words would have been needed to eliminate alternative explanations.

1.6 A study in three German schools

- 230 pupils were involved in the study across three schools in Germany. All pupils were in class 5 (10 to 11 years of age). All pupils had transferred from primary schools to the secondary schools when the study was begun.
- The pupils used phase-6 to different extents, but all were learning English as a foreign language. The study involved the use of pre- and post-test papers to identify levels of recognition of words and phrases in German and English. Test paper evidence was gathered by a member of the research team, in conjunction with members of phase-6 and selected schools. The test papers were available to the research team.
- This study provided findings not just about impacts of phase-6, but also about different patterns of use by different pupils in different schools, ways to use phase-6 software that might support effective language learning, ways that phase-6 might be developed further in the future, and importantly, indications about school differences and teacher approaches that could be major factors masking identification of some impacts.
- Pupils who had used phase-6 in their previous primary schools saw it as helpful and tended to use it again.

- phase-6 was quite usable. When introduced to new study tools, pupils often believe that they will use them, but then fail to do so. For phase-6, pupils' good intentions at the beginning of the year corresponded significantly with a greater likelihood of them actually using phase-6 during the year.
- Girls translated significantly more of the items correctly from Paper 1 than did boys. However, the effect of gender was not significant at the end of the test period.
- Pupils were significantly better at receptive translations than productive ones in both papers (and certainly this result, taking research into language translation into account, would be expected). However, this may be an important conclusion for phase-6 users, in terms of the balance of access that pupils have to receptive and productive vocabulary.
- Pupils translated items correctly significantly more often in Paper 1 and in Paper 2 when a sentence was present for context than when it was not. Context provided a benefit for verbs and adjectives, but not for nouns and preposition or connectives in Paper 1, while in Paper 2 context provided little benefit for nouns, but improved translations for the other parts of speech. This result may be important in terms of the forms of questions within vocabulary training packages when phase-6 is used.
- Nouns were more often translated correctly in Paper 1 than the other parts of speech. This pattern is usual and is often attributed to the more concrete nature of many nouns. In Paper 2, prepositions and connectives were translated correctly almost as often as were nouns. The translation of verbs did not follow a pattern that was expected or in common with other forms of speech. This is an area that should be further explored, particularly as the verb forms used were infinitives rather than conjugated forms being involved.
- School factors had a significant effect on pupils' initial performance and later performance. Based on large differences between schools, it would appear that other aspects of English language learning may well influence pupils' learning more than has their use of phase-6. Evidence from one school indicated that a wide variety of different retrieval and practice approaches were adopted by pupils, but that the most common were concerned with using word lists in textbooks (with someone testing this list or it being written out in a vocabulary book). The emphasis here is on the testing of vocabulary within short time periods. phase-6 uses an increasing time interval and is concerned with long-term memorisation. Teachers and pupils have focused in this school more on the adoption of short-term approaches. If pupils are to memorise effectively, to meet the needs of both short and long term learning, it is likely to be important that both forms of practice are introduced and balanced.
- Two analyses run at a specific school level suggest that phase-6 is supporting certain groups of pupils. In one school, girls who gained high marks in Paper 1 gained marks that were well above the average in Paper 2 when they also used phase-6. It is possible that pupils who score high marks use techniques that do not involve high levels of social interaction. Hence, use of phase-6 can match the approaches to learning taken by this group of pupils. Their independent and persistent use of phase-6 could allow them to explore vocabulary learning so that they can gain higher marks than their peers who use phase-6 less persistently. In another school, pupils who used phase-6 daily gained higher test scores, when translating nouns, adjectives, prepositions and connectives in sentence contexts in a receptive direction, at levels of statistical significance when compared to those using phase-6 less regularly.

1.7 Conclusions in terms of future directions and uses

- Findings suggest that teachers and learners might well gain from using phase-6 more effectively if it was used in certain ways. Some initial guidance points for learners and teachers arise from the findings of this study.
- One important point that learners and teachers should consider is the time interval set between phases. The time interval should match with an individual's initial level of understanding of subject content, and approaches to learning and memorisation. phase-6 provides a facility for the user to alter the time intervals between phases. In practice this facility has not been seen in use a great deal. Teachers should consider how time intervals between phases might support their groups of learners most effectively, perhaps suggesting time intervals based on the notion of fast learners,

medium paced learners, and learners who take more time in their learning. In the future an early test within the system might support this differentiation to a greater extent, offering some levels of objective indicators.

1.8 Recommendations for further studies focused on identification of learning impacts

- From a research point of view, phase-6 provides an opportunity to set up focused research studies, since the resource offers support in quite specific areas of learning (in the areas of memorisation, retrieval and practice). Hence, it is possible for studies to be set up that explore impacts in these specific learning areas. Specific test items that identify levels of memorisation, retrieval and practice can be selected to provide appropriate measures of impact.
- However, it is clear from findings of the studies reported here that teacher approaches can dramatically influence impacts of the phase-6 resources. Consequently, adequate controls and careful matching of samples are needed with future research studies to ensure that the precise influence of phase-6 can be rigorously identified.
- The research undertaken provides useful indicators of how an ideal study might be set up to gather robust and rigorous evidence about impacts on learning aspects concerned with memorisation, retrieval and practice. Guidance and principles reported by the National Mathematics Advisory Panel in the US (accessible at http://www.ed.gov/about/bdscomm/list/mathpanel/reports.html) regarding standards of evidence for influencing educational decisions are particularly useful and relevant in this context.
- A future study will need to gather both qualitative and quantitative evidence. These forms of evidence will allow levels of impact to be quantified, as well as reasons why.
- Use of phase-6 facilities is most developed currently in terms of language learning. It would seem appropriate in a future study to focus on the impact of phase-6 on language learning in year 9, and its potential impact on subsequent choice of subject at GCSE.
- The study will need to generate pre- and post-test data on the learning of words or other subject matter, both using phase-6 and without phase-6. The best complete design should involve pairs of schools that are well matched in terms of their student intake and their teaching materials and methods, so that test data can be gathered independently (using phase-6 in one school and not using phase-6 in the paired school). However, in order to remove the possibility of impact of a Hawthorn effect, it will be important that the study is set up so that data is gathered without influence of one school on the other, so that the school not using phase-6 is unaware of this practice if at all possible. In any pair of schools, teacher and learner backgrounds, methods and approaches should be similar, other than the use of phase-6. It will be possible to compare the two groups on pre-tests to show general equivalence.
- For wider generalisation of results, it is important that paired schools cover different geographical areas, different locality settings (urban, rural and suburban settings), socio-economic settings, and banded, streamed or mixed groupings used for class teaching. Cohorts of some 100 students in each school would be ideal in terms of gathering sufficient levels of data.
- So far, the studies have focused on the memorisation of words and phrases in text form. phase-6 will support multi-modal study. Although this aspect has not been investigated as a part of the study reported here, it is expected that multi-modal functionality will be beneficial in the long term.
- Multi-modal online access is likely to make a useful contribution to the research, since the ability to gather data in written as well as recorded form would be clearly advantageous. This would be especially useful if students were able to record and review their spoken responses against exact spoken responses.

2. BACKGROUND

2.1 Evaluating impacts of educational technologies on learning

Over the past 20 years and more, there have been ranges of technology-based resources that have been developed and implemented in practice both in schools and outside schools to support learning. When technology-based resources were being developed some 20 years ago, the level of technology equipment available and accessible inside and outside schools was low. The implementation of these early resources were sometimes accompanied by research studies that sought to identify the impact of the technology-based resources on learning, often by comparing test outcomes with a non-intervention or non-technological intervention, in a parallel control group. A number of these controlled studies were set up, more often in the US than in the UK, as technology-based resources were developed earlier within the US than they were in the UK. Some of those resources, such as large-scale integrated learning systems, were introduced into the UK, and independent studies were conducted to look at their impacts (reported in NCET, 1994; NCET, 1996; Wood, 1998).

Since those studies were conducted, there has been an enormous increase in the ranges and levels of technology that have become accessible to learners. The greatest increase in levels of ICT resources and access to ICT in schools in England has occurred since 1998, implemented through two successive major national policies: the National Grid for Learning (NGfL) initiative; and the ICT in Schools initiative. Over this same period, there have been increased resources and access to resources available across countries widely, and widely within homes. The more recent NGfL and ICT in Schools initiatives in England have been accompanied by a series of evaluation and research studies, exploring ongoing implementation and outcomes (with reports from, for example, Harrison *et al*., 2002; Somekh *et al*., 2001; 2002a; 2002b; Cox *et al.*, 2003a and 2003b; Pittard *et al*., 2003; Passey and Rogers, 2004; Underwood *et al*., 2005). However, the nature of these research studies has been different, since the identification of specific outcomes from specific technologies has been made increasingly difficult. When a range of technology-based resources are present, it is much more difficult to identify the impacts of specific resources and technologies. The body of research examining the impact of ICT on learning, learners, teaching and education, undertaken over the past 10 years, therefore, offers a fundamental level of understanding at a more generalist level, providing wide and general indicators of outcomes and impacts (especially when national studies, notably Becta, 2001a; 2001b; 2003a; 2003b; and Ofsted, 2001; 2002; 2004, looking at overall subject attainment related to levels of technology-based resources and how they are managed are considered). Some research and evaluation studies have reported that ICT can have an impact upon learning when that learning is measured by subject attainment. Most notably, perhaps, Harrison *et al*. (2002, p. 2), found that:

> "A statistically significant positive association between ICT and National Tests for English was found at Key Stage 2. Positive associations were also found for mathematics at Key Stage 2, although they were not as striking and not statistically significant. … A statistically significant positive association between ICT and National Tests for science was found at Key Stage 3, but there were no other clear-cut associations at Key Stage 3. … At Key Stage 4, there was a statistically significant positive association between ICT and GCSE science and in GCSE design and technology."

More recent work that has looked at qualitative impacts of specific technologies on learning has begun to identify more exactly those specific learning processes where impacts might arise for specific technologies. This suggests that more specific studies should be undertaken (argued in Passey, 2006), in order to explore the impacts at levels where more precise guidance and advice on uses will support both learners and teachers. The most recent UK-based quantitative studies looking at technology-based resources has been the series of Fischer Family Trust reports on impacts of SAM Learning online (often outside school) revision materials on GCSE results (Fischer Family Trust, 2003; 2004). The exploration of learning impacts arising from uses of phase-6 is an example of a quantitative study that is feasible since specificity of use enables potential distinction of its individual impact. phase-6 offers learning support in terms of specific learning processes; the aims of the resource are stated in terms of impacting on revision, memorisation and recall. It is possible for studies to be set up that can explore impacts in these learning areas.

However, setting up studies for a technology-based resource that aims to impact on certain learning processes still requires adequate elimination of possible 'noise' (that is, those factors that can impact on both uses of the resource, and uses of the test items involved). Findings from existing studies point towards a range of factors that can lead to 'noise', and that can contribute to outcomes. Studies where pupils have used ICT on its own (that is, without teacher intervention or support) have rarely identified an enhancement of attainment beyond an initial period of time (teachers and observers have reported a matter of a few months with some forms of ILS, for example). Becta (2001b, p. 8) indicated the role that teaching practices might play in terms of contributory factors:

> "Analysis of the Ofsted data on quality of ICT use reveals that attainment is even higher when high levels of ICT resource are combined with 'Good' ICT teaching. On average 69% of pupils in schools with 'Very good' ICT resources attained at least five GCSEs. When 'Very good' resources are combined with 'Good' ICT teaching, this proportion rises to 72%."

As Cox *et al*. (2003a, p. 3) stated:

> "There is a strong relationship between the ways in which ICT has been used and pupils' attainment. This suggests that the crucial component in the appropriate selection and use of ICT within education is the teacher and his or her pedagogical approaches. Specific uses of ICT have a positive effect on pupils' learning where the use is closely related to learning objectives."

It is clear from this present study that teacher approaches can dramatically influence impacts of the phase-6 resources. For example, in one school, some teachers use regular vocabulary tests to encourage pupils to revise and remember words. These vocabulary tests are undertaken at intervals that do not correspond well to the time intervals used within phase-6 for the memorisation of words. Pupils need to use phase-6 for some months before words will be placed into the sixth box, but teachers are testing pupils on these words within a matter of days of weeks, rather than months. The practice of using regular vocabulary tests, therefore, tends to push pupils towards using other forms of revision and memorisation. Further than this, pupils then see the value of those short-term methods, and may well not be encouraged to use longer-term methods such as uses of phase-6 to revise and remember words.

This stage of the research will enable an identification of factors and features that impact upon uses of the phase-6 resources. In the light of this knowledge, it will be possible to indicate the range of ways that a study would need to be set up in order to have rigorous and robustness with regard to findings. For findings to be accepted and used, it is vitally important that they are seen as having been carried out with rigour, and national guidance on this matter indicates how teachers are being encouraged to view evidence arising from resource evaluation studies. The US Department of Education, Institute of Education Sciences, and National Center for Education Evaluation and Regional Assistance (2003) published a guidance document that advised teachers and policy makers to consider the need for rigorous evidence to support practice. They offered advice about:

- Randomised controlled trials.
- Evaluating whether an intervention is backed by 'strong' evidence of effectiveness.
- Evaluating whether an intervention is backed by 'possible' evidence of effectiveness.
- Important factors to consider when implementing evidence-based interventions in schools or classrooms.

Further guidance is available from the report on standards of evidence (Reyna, Benbow, Boykin, Whitehurst and Flawn, 2008), which distinguishes both on the basis of the quality of the research design and the setting of the research (classroom or laboratory) as a guide to generalisability.

The aim of this stage of the research study for phase-6 is to consider how evidence would be provided of the forms indicated within guidance, so that its robustness and rigorousness can be used to critically inform teachers and policy makers not only within the UK, but more widely. Software and resources within the UK have been developed on the basis of need; learning and teaching needs identified by practitioners or researchers have led to the developments of software or resources to attempt to address

these needs. phase-6 is, within the UK context, perhaps unique, in that it is a technology-based resource that has been developed on the basis of learning theory. From an evaluation and research point of view, therefore, aims of studies can be more critically focused than might be the case with other software or resources.

2.2 Theoretical background

The phase-6 technique is related to the retrieval practice approach to improving memory. In terms of applications and practices arising, it is possible that one line of thought developed in mainland Europe while the other developed in North America and Britain. Both are based on Ebbinghaus's early evidence that practice testing benefits learning and that spaced practice is more beneficial than massed practice.

In mainland Europe pupils in schools (starting with primary schools) use boxes, divided into five spaces, to support the revision and memorisation of words or phrases. This box, described by Leitner (originally in 1972, within the fourteenth edition in 1995), has spaces that are increasingly longer, enabling an accumulation of increasing numbers of words into the higher spaces. The concept described by Leitner, as used by pupils to this day, suggests that cards are used for self-testing, and that those in the first space are worked on until they fill the second space. When the second space is full, they are worked on until they fill the third space, and so on. Leitner does not specify the time interval between test spaces.

phase-6 has taken this concept, and has changed the space or distance perspective to a time perspective, so that cards are tested, and when moved to a new space, there is a specific time interval before they are tested again. Certainly the idea that there is memory loss over time, and the idea that regular testing can benefit memorisation, are reported by Leitner (although the Ebbinghaus memory loss curve shown in his book indicates that most memory loss occurs within a day).

In recent years there has been considerable research in the US and the UK on the benefits of spacing retrieval practice, especially with expanding intervals between retrieval attempts (for example, Landauer and Bjork, 1978; Roediger and Karpicke, 2006; Morris and Fritz, 2007). Expanding retrieval practice has been shown to be effective in many situations for normal adults, pre-school and older children, and the elderly (shown in, for example, Fritz, Morris, Acton, Voelkel and Etkind, 2007; Morris, Fritz, Jackson, Nichol and Roberts, 2005; Fritz, Morris, Nolan and Singleton, 2007). The technique involves attempting to retrieve items on a few occasions; for expanding practice, retrieval is cued initially soon after the information is first encountered, and then following increasing delays.

Expanding schedules have been demonstrated to be better than fixed schedules when feedback (giving the correct answer) does not follow the retrieval attempt (reported by Cull, Shaughnessy and Zechmeister, 1996; Fritz and Morris, 2003; Landauer and Bjork, 1978). However, when re-presentation of the correct information is provided as feedback, fixed spacing is as beneficial as expanding spacing (Cull, 2000; Fritz and Morris, 2003; Roediger and Karpicke, 2006).

A number of theoretical explanations of expanding retrieval practice have been offered, including: the need for desirable difficulty in the learning task (Bjork and Bjork, 1992) and encoding variability (Neuschatz, Preston, Toglia and Preston, 2003).

2.3 Context

phase-6 AG is a Swiss-based company that has produced an innovative piece of software to support aspects of learning where memorisation is a focal need. phase-6 is the name given to a learning resource, created in software form, devised to support learners with memorisation (of facts, definitions, or phrases, for example). The development of the software has been underpinned by an accepted theoretical framework, which describes how memorisation can be acquired and enhanced through five successive periods of repetition and revision, increasing in length over time (called phases, according to the work of Ebbinghaus). The theoretical framework of Ebbinghaus has been used to develop a widely used practical technique (devised originally as a system of boxes with cards

that could be moved according to the phase of memory reached, and used commonly by pupils in German schools). The purpose of this system, and that of the software system, is to support the memorisation of words, events, definitions or phrases through regular, but increasingly protracted, revision phases. This practically based system has now been created in an ICT form. The phase-6 software facilities are recognised by the company as being used increasingly by pupils in Germany.

2.4 Initial Research Studies and Approaches

phase-6 has commissioned the research group at Lancaster University to undertake initial studies, in order to look at potential impacts of its software. Two exploratory studies were established, and data from these studies was made accessible to the research group in August 2008.

The two studies involved:

- Two hundred and thirty pupils across three schools in Germany, using the phase-6 software to different extents, but all concerned with the learning of English as a foreign language, and with the learning of vocabulary associated with the language learning needs of Class 5 (10 to 11 year old) pupils. The identification of potential impact has so far involved a scoping activity in a small number of German schools, and the development and use of pre- and post-test papers, to identify levels of recognition of words and phrases in German and English, appropriate to the needs of this age range of pupils. Preliminary work to gather test paper evidence was undertaken by a member of the research group, in conjunction with members of phase-6 and selected schools.
- One hundred and seventy-two pupils in a school in California, using phase-6 to support the learning of Spanish language vocabulary, using phase-6 for some words, but not others. This approach was taken in order to provide baseline-learning measures (control data) for each individual student. This preliminary work was undertaken by phase-6 in California, in conjunction with a selected school.

phase-6 AG commissioned the research group to undertake:

- The analysis of the results from the tests undertaken by pupils in German schools.
- The identification of findings arising from those results, and the identification of any further data needed to support and enhance the findings further.
- The reporting of the results of the findings from phase-6 use by the sample of pupils in German schools for wider dissemination.
- The review of the data collected from the tests undertaken by pupils in the school in California.
- The identification of findings arising from those results, and the identification of any further data needed to support and enhance the findings further.
- The reporting of the results of the findings from phase-6 use by the sample of pupils in the school in California for wider dissemination.

3. THE RESEARCH STUDY IN GERMAN SCHOOLS

3.1 Identifying the focus for the study

In May 2007, visits to six schools in Germany enabled evidence from teachers and a range of pupils across year groups, to be gathered about their uses of phase-6 within the schools. The interviews indicated that the majority of uses in the schools were to support the teaching of English. It was decided at that time that a study should, therefore, focus on the identification of impact on English language learning. Class 5 pupils, 10 to 11 years of age, were selected for the study. This year group is the first year group in secondary schools, so their background uses of phase-6 in primary schools would be likely to be patchy at best.

3.2 The study approach

In order to gain some consistency across schools, it was decided that a test paper would be devised so that evidence about individual pupil memorisation of words and phrases in English could be gathered at two points across a school year. A test paper was created, and this test was administered in September 2007 (at the beginning of the school year for the new Class 5 pupils) and in April 2008 (after some 6 months of English language teaching). The use of phase-6 is a personal pupil's choice, guided by teachers. The test papers gathered self-reports about pupils' prior use of phase-6 in primary schools, and their uses of phase-6 between September 2007 and April 2008.

3.3 The test papers

The test papers were in German, so that pupils were using their native language to read questions and instructions. Each paper consisted of two sections. The first section gathered background information, shown in the table following.

Paper 1
a. phase-6 was used in primary school
b. The pupil expects to use phase-6 this year
c. The pupil expects phase-6 to help with learning English

Paper 2
a. The pupil worked with phase-6 this year
b. The pupil used phase-6 almost every day
c. The pupil used phase-6 once or twice a week

Table 1: Background details gathered on test papers

The second section was identical in both test papers. The second section was divided into nine parts, covering different forms of words, both within a sentence context and outside a sentence context. Each of the nine parts listed six words or phrases, and pupils needed to write into an appropriate box the German or English equivalent of those words. In each part there were three words or phrases in German (for pupils to give the English equivalent) and three words or phrases in English (for pupils to give the German equivalent); translation direction alternated from one test item to the next. The nine parts covered word and phrase forms shown in the table following.

Word or phrase form
a. Nouns (as single words)
b. Verbs (in sentence context)
c. Adjectives (as single words)
d. Prepositions and connectives (in sentence context)
e. Phrases
f. Prepositions and connectives (as single word)
g. Adjectives (in sentence context)
h. Verbs (as single words)
i. Nouns (in sentence context)

Table 2: Word and phrase forms covered in each test paper

In German schools, there are mainly three text books used to support the teaching of English, published by Cornelsen, Klett, and Diesterweg. Each of the text books list words and phrases that pupils should know by the time they have completed the Class 5 year. Although there is some variation across textbooks, words and phrases were selected for the test papers that were common to all three textbooks.

3.4 Coding of the paper responses

In July 2008, all test papers completed by pupils in September 2007 and April 2008 were marked and coded, and the results were collated into a MS Excel spreadsheet. School name, pupil name, pupil gender, and the test paper number were recorded in all cases. Responses from pupils to the questions in the first part of each paper were recorded by using a 0 to show a 'no' response, and a 1 to show 'yes' response. Responses from pupils to the questions in the second part of each paper were recorded by using a 0 to show an incorrect or null response, a 1 to show a correct response, and a 2 to show a correct but synonymous response (not shown in the marking scheme). These responses were later recoded to allow independent analysis of In the second part, correct and incorrect spellings were also recorded. The data are from all pupils in Class 5 in three schools.

3.5 Structure of the data within the spreadsheet

The MS Excel spreadsheet used to collate the test data contained data in six worksheets: all responses for both papers (n=555); all responses to paper 1 (n=291); all responses to paper 2 (n=263); all paired responses (n=460); all paired responses to paper 1 (n=230); and all paired responses to paper 2 (n=230). Each worksheet was set out similarly, showing school name, pupil name, pupil gender, paper number, responses to questions in part 1, and responses to questions in part 2 (in terms of both correct or incorrect responses, and correct or incorrect spelling).

3.6 Questions for analysis of the responses

A number of key questions were identified, that could be posed of the data in the worksheets.

For all results in paper 1:

- Is the balance of girls and boys about the same?
- What is the overall level of correct response for each form of word or phrase (nouns, verbs, adjectives, prepositions and connectives, and phrases)?
- Is it the same for each school?
- What is the overall level of correct spelling for each form of word or phrase (nouns, verbs, adjectives, prepositions and connectives, and phrases)?
- Is it the same for each school?
- Does having words set in sentence contexts appear to make any difference to response levels?
- Is it the same for each school?
- Does prior use of phase-6 seem to make any difference?
- Does expected use of phase-6 seem to make any difference?
- Does expected learning from use of phase-6 make any difference?

For all results in paper 2:

- Is the balance of girls and boys about the same?
- What is the overall level of correct response for each form of word or phrase (nouns, verbs, adjectives, prepositions and connectives, and phrases)?
- Is it the same for each school?
- What is the overall level of correct spelling for each form of word or phrase (nouns, verbs, adjectives, prepositions and connectives, and phrases)?
- Is it the same for each school?
- Does having words set in sentence contexts appear to make any difference to response levels?
- Is it the same for each school?
- Does use of phase-6 seem to make any difference?
- Does high level of use of phase-6 seem to make any difference?
- Does some level of use of phase-6 make any difference?

For all paired results in papers 1 and 2:

- Is the balance of girls and boys about the same?
- What is the overall level of correct response for each form of word or phrase (nouns, verbs, adjectives, prepositions and connectives, and phrases)?
- Is it the same for each school?
- What is the overall level of correct spelling for each form of word or phrase (nouns, verbs, adjectives, prepositions and connectives, and phrases)?
- Is it the same for each school?
- Does having words set in sentence contexts appear to make any difference to response levels?
- Is it the same for each school?
- Does no use of phase-6 seem to make any difference?
- Does high level of use of phase-6 seem to make any difference?
- Does some level of use of phase-6 make any difference?
- Is there any evidence that phase-6 is making any difference identified (or could other factors such as conscientiousness account for any difference identified)?

4. FINDINGS FROM THE STUDY IN GERMAN SCHOOLS

4.1 An overview of test data from the three German schools

The table following gives totals for data fields from all three German schools.

Total number of pupils	230
Phase-6 users in previous primary schools	25
Phase-6 users in the secondary schools	114
Every day users	48
Once or twice a week users	37
Paper 1 correct answers	3741
Average number of correct answers in Paper 1	16
Paper 2 correct answers	7996
Average number of correct answers in Paper 2	35
Total difference between Papers 1 and 2 correct answers	4255
Average number of improved responses from Paper 1 to Paper 2	19

Table 3: Overview of responses in both test papers (N = 230)

Pearson's correlation tests were run on levels of phase-6 uses against test results for each paper and their difference, for all data from the three German schools. The results are shown in the table following. For all correlations, N = 230

		Total correct Paper 1	Total correct Paper 2	Difference
Primary user	Pearson Correlation	.095	-.063	-.137(*)
	Sig. (2-tailed)	.151	.342	.038
phase-6 user	Pearson Correlation	.036	.043	.010
	Sig. (2-tailed)	.589	.521	.878
Every day user	Pearson Correlation	.052	.029	-.016
	Sig. (2-tailed)	.429	.658	.814
Once or twice a week	Pearson Correlation	-.059	-.031	.019
	Sig. (2-tailed)	.376	.637	.774

Table 4: Results of correlation tests between level of phase-6 user and total correct responses
(Note: ** Correlation is significant at the 0.01 level (2-tailed). * Correlation is significant at the 0.05 level (2-tailed).)

These tests have not revealed any highly significant correlations between results and background usage levels. It is possible, therefore, that the impact of phase-6 is more specific, perhaps impacting upon specific groups of pupils; there appears to be no simple relationship between level of use, test result and improvement level. It should be noted that pupils involved in this study have not used phase-6 alone for supporting revision and memorisation of words. In school 2, for example, teachers of two of the classes asked pupils about the methods they used. The results are shown in the table following.

Revision methods used	**Frequency**
phase-6 only	5
phase-6 and handwritten lists of words in a vocabulary book	4
phase-6 and handwritten lists of words in a vocabulary book and someone testing you	3
phase-6 and word lists in the textbook	7
Box with cards alone	2
Box with cards and handwritten lists of words in a vocabulary book	2
Handwritten lists of words in a vocabulary book and someone testing you	2
Word lists in the textbook and someone testing you	10
Word lists in the textbook and handwritten lists of words in a vocabulary book	13
Total	48

Table 5: Levels of different revision methods used by pupils in 2 classes in one of the schools

It is clear from this evidence that many pupils use alternative methods, and that less than half the pupils use phase-6. It is also clear from the evidence gathered by the two teachers that methods are not used in the same proportion in each class. For example, in one class a large number of pupils (13 in total) use word lists in the textbook and handwritten lists of words in a vocabulary book, while no pupils use this mixture in the other class. The reason for the range of variation is clear when teachers explain that they use short vocabulary tests in class, very frequently, to encourage the learning and memorisation of words. Teachers, therefore, encourage pupils to use techniques that will support memorisation in the short-term, such as uses of handwritten lists, and testing one another verbally. Phase-6, because of the time intervals between revision testing, supports a much longer-term learning approach. Clearly, if teachers do not encourage this longer-term learning approach, either instead of or as well as short-term memorisation techniques, then outcomes are not likely to be strong.

4.2 Results at a school level

Three schools, including 326 pupils in total, wrote answers to at least one paper. The numbers of pupils from each school answering each paper, and the numbers who answered both papers, are summarised in the table following. Overall, the difference between the number of girls and boys was not statistically significant, $\chi^2(1, N = 230) = 2.10$, $p = .15$.

School	Gender	Paper 1	Paper 2	Both papers
School 1				
	Girls	37	39	30
	Boys	48	49	40
	Total	87 *	88	70
School 2				
	Girls	52	53	51
	Boys	53	50	44
	Total	105	103	95
School 3				
	Girls	63	50	45
	Boys	37	23	20
	Total	100	73	65
All schools				
	Girls	152	142	126
	Boys	138	122	104
	Total	292	264	230

* Two pupils were multiply coded (G/B, B/B)

Table 6: Total numbers of pupils completing each test paper by school and gender

It should be noted that some abnormalities identified on inspection of the initial spreadsheet were corrected. One of the pupils from School 3 was listed as a boy on the Paper 1 list and a girl on the Paper 2 list. The spreadsheet was corrected so that the pupil appeared as a girl on both lists. In School 1, a pupil identified by one first name sat Paper 1 but was identified with another first name when sitting Paper 2. These were matched as a pair; the misreading of the pupil's written first name gave rise to this difference.

In paper 1 at the beginning of the year pupils were asked about their previous experience with phase-6 and about their plans and expectations. In paper 2 at the end of the year pupils were asked about their actual use of phase-6 during the year. The number of pupils who replied 'Yes' to the questions are reported in the table following.

School / Gender	*N*	Beginning of the year: Previous use	Beginning of the year: Plan to use it	Beginning of the year: Expect it to help	End of the year: Used it	End of the year: … almost daily	End of the year: … at least weekly
School 1							
Girls	30	4	17	21	16	5	4
Boys	40	3	31	34	23	13	8
Total	70	7	48	55	39	18	12
School 2							
Girls	51	4	40	40	27	14	4
Boys	44	6	34	33	19	7	7
Total	95	10	74	73	46	21	11
School 3							
Girls	45	5	29	37	23	6	8
Boys	20	3	8	14	9	3	5
Total	65	8	37	51	32	9	13
All schools							
Girls	126	13	86	98	66	25	16
Boys	104	12	73	81	51	23	20
Total	230	25	159	179	117	48	36

Table 7: Previous pupil experience and plans for use of phase-6 by school and gender

Note: The spreadsheet included 4 pupils who did not report using phase-6, but then reported having used it either almost daily or at least weekly. These instances were coded as having used phase-6.

Only 25 pupils had previous experience of phase-6; 21 of them expected that it would help them to learn English, 19 planned to use it, and at the end of the year 17 of these 25 students reported that they had used it, all of these daily or weekly. **Although the numbers are small, these data suggest that once students have used phase-6 they see it as helpful and tend to use it again.**

Of the 159 pupils who expected to use phase-6 during the year:

- 148 expected that phase-6 would help their learning of English at the end of the year.
- 104 (65%) reported that they had used phase-6,
  - 43 reported using phase-6 almost every day.
  - 33 at least once or twice each week.

Of the 71 pupils who did not expect to use phase-6 during the year:

- 31 expected that phase-6 would help their learning of English.
  - 7 later reported having used phase-6: 2 daily and 1 weekly at the end of the year.
- 13 (18%) reported that they had used phase-6
  - 5 almost every day.
  - 3 at least weekly.

Expectations and use of phase-6
Number of pupils:

		Used: Y	Used: N	Total
Expected to use	Y	104	55	159
	N	13	58	71
Total		114	116	230

Table 8: Numbers of pupils expecting to use phase-6 compared to actual use

Thus pupils' good intentions corresponded significantly with a greater likelihood of actually using phase-6 during the year, $\chi^2(1, N = 230) = 43.57$, $p < .001$.

4.3 Forms of questions within the test papers

The test taken by the pupils at the beginning and end of the year included 54 translations: 27 from English to German (receptive vocabulary) and 27 from German to English (productive vocabulary). In learning foreign language vocabulary, pupils learn receptive translations (translating foreign to native language) much faster than (and perhaps differently from) productive translations (translating native to foreign).

There were three items of each type for each of nine translation tasks: Eight of the tasks required pupils to translate nouns, verbs, adjectives, and prepositions/connectives either presented in isolation or in the context of a sentence; the ninth task was translation of phrases. The six translations for each task were blocked and appeared in the following order: noun-no context, verb-sentence, adjective-no context, preposition/connective-sentence, phrase, preposition/connective-no context, adjective-sentence, verb-no context, and noun-sentence. Within each block receptive and productive translations alternated.

Translations were scored in two ways, allocating one mark for each correct translation: The strict score was based on pupils producing anticipated translations and the more flexible score allowed for correct translations that were not anticipated (and therefore not part of the study materials). A score for correct spelling was also calculated, allocating one mark for each correct spelling. In the data reported below, spelling marks were only awarded for correct translations; incorrect translations that were nevertheless correctly spelled are not included. The latter situation arose for just 12 individual English to German translations and 11 German to English translations, distributed across 8 pupils. Four pupils each contributed just one correctly spelled incorrect translation and the other four pupils contributed 3, 4, 5, and 7 correct spellings of incorrect translations.

4.4 Performance on the first test

The translation scores from the first test – at the beginning of the year – are summarised in the table following. Means are reported (with standard deviations in parentheses). The maximum possible value for each cell is 3.

	Receptive - English to German				Productive - German to English			
Word type	No context		Sentence context		No context		Sentence context	
Strict scores								
Nouns	1.0	*(0.79)*	1.5	*(0.70)*	1.3	*(0.55)*	0.7	*(0.80)*
Verbs	0.6	*(0.85)*	0.6	*(0.64)*	0.6	*(0.84)*	1.0	*(0.56)*
Adjectives	0.6	*(0.64)*	1.4	*(0.74)*	0.3	*(0.64)*	0.7	*(0.70)*
Prepositions and connectives	0.9	*(0.76)*	0.9	*(0.95)*	0.7	*(0.69)*	0.7	*(0.77)*
Phrases	0.3	*(0.60)*			0.3	*(0.56)*		
Flexible scores								
Nouns	1.0	*(0.79)*	1.5	*(0.70)*	1.3	*(0.55)*	0.7	*(0.80)*
Verbs	0.6	*(0.85)*	0.6	*(0.64)*	0.6	*(0.84)*	1.0	*(0.56)*
Adjectives	0.7	*(0.71)*	1.5	*(0.78)*	0.3	*(0.64)*	0.7	*(0.70)*
Prepositions and connectives	1.1	*(0.86)*	0.9	*(0.95)*	0.7	*(0.70)*	0.7	*(0.77)*
Phrases	1.1	*(0.45)*			0.3	*(0.56)*		

Table 9: Translation scores for word types compared to direction and context

In order to detect interactions as well as main effects, school, gender, translation direction (receptive or productive), sentence context, and part of speech (noun, verb, adjective, preposition/connective) were investigated with a five-way mixed ANOVA for pupils' translations at the beginning of the year. Translation of phrases was not included in this analysis.

School: School had a significant effect on pupils' initial performance, for strict scoring $F(2,224) = 8.89$, $MSE = 2.62$, $p < .001$, partial $\eta^2 = .07$; for flexible scoring $F(2,224) = 8.78$, $MSE = 2.87$, $p < .001$, partial $\eta^2 = .07$. Tukey posthoc comparisons show that the pupils at School 3 produced more correct translations than did the pupils at School 2, $p = .001$. The latter produced more correct translations than did pupils at School 1, $p = .014$ for strict and .015 for flexible scoring. School 3's pupils also produced significantly more translations than School 1's, $p < .001$. Later in the report it is noted that school is also related to phase 6 usage and to translation performance at the end of the year.

	Strict		Flexible	
School	Mean	SEM	Mean	SEM
School 1	0.67	0.049	0.67	0.051
School 2	0.83	0.042	0.84	0.044
School 3	0.96	0.054	0.99	0.057

The maximum possible score was 3 for each cell.
Table 10: School scores for Paper 1

Gender: Girls translated significantly more of these items correctly than did boys. For strict scoring, $F(1,224) = 23.11$, $MSE = 2.62$, $p < .001$, partial $\eta^2 = .09$; for flexible scoring, $F(1,224) = 22.35$, $MSE = 2.87$, $p < .001$, partial $\eta^2 = .09$.

	Strict		Flexible	
Gender	Mean	SEM	Mean	SEM
Girls	0.95	0.037	0.97	0.039
Boys	0.68	0.042	0.69	0.044

The maximum score was 3 for each cell.
Table 11: Paper 1 scores by gender

Translation direction: As is usual when learning foreign vocabulary, pupils were significantly better at receptive translations than productive ones; for strict scoring $F(1,224) = 90.31$, $MSE = 0.30$, $p < .001$, partial $\eta^2 = .29$ describing a very large effect, for flexible scoring $F(1,224) = 113.31$, $MSE = 0.33$, $p < .001$, partial $\eta^2 = .34$ describing a very large effect.

The direction of the translation did not interact individually with either school or gender. The three way interaction was statistically significant, but the effect was small and perhaps not worthy of interpretation. For strict scoring $F(2,224) = 3.06$, $MSE = 0.303$, $p = .049$, partial $\eta^2 = .03$; for flexible scoring $F(2,224) = 3.73$, $MSE = 0.332$, $p = .025$, partial $\eta^2 = .03$.

	Strict		Flexible	
Direction	Mean	SEM	Mean	SEM
Receptive	0.91	0.031	0.94	0.034
Productive	0.72	0.028	0.73	0.028

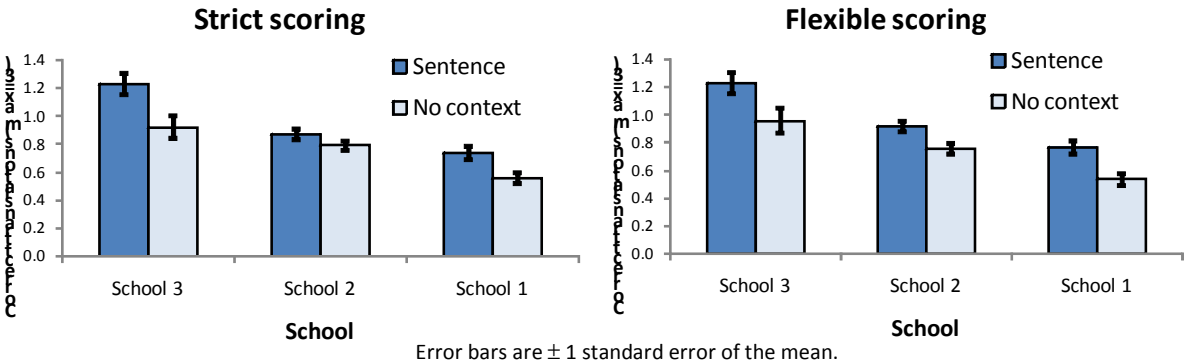
The maximum possible score was 3 for each cell.
Table 12: Paper 1 scores by translation direction

Context: Pupils translated items correctly significantly more often when a sentence was present for context than when it was not; for strict scoring, $F(1,224) = 107.13$, $MSE = 0.30$, $p < .001$, partial $\eta^2 = .32$ describing a very large effect; for flexible scoring, $F(1,224) = 80.96$, $MSE = 0.30$, $p < .001$, partial $\eta^2 = .27$ describing a very large effect. This effect might be due to the support provided by the context or to having more familiar or easier words offered with a sentence for context; further research could vary the test among the pupils so that the same words were offered with sentences for some students and without them for others.

	Strict		Flexible	
Context	Mean	SEM	Mean	SEM
Sentence	0.91	0.030	0.92	0.030
None	0.72	0.029	0.75	0.031

The maximum possible score was 3 for each cell.
Table 13: Paper 1 scores by sentence context

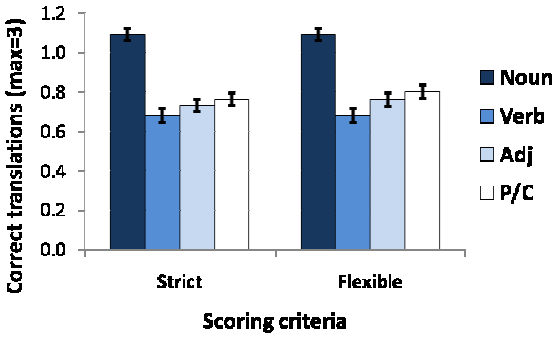
Context interacted significantly with school, but not with gender and the three-way interaction was not significant. For the context x school interaction for strict scoring, $F(2,224) = 13.44$, $MSE = 0.30$, $p < .001$, partial $\eta^2 = .11$; for flexible scoring, $F(2,224) = 13.87$, $MSE = 0.20$, $p < .001$, partial $\eta^2 = .11$. The figure below illustrates the nature of the interaction: For pupils at School 2 the sentence context provided far less benefit, especially for strict scoring.



Error bars are ± 1 standard error of the mean.

Figure 1: Effect of sentence context by school

Part of speech: The part of speech played a significant role in initial test translations, for strict scoring, $F(3,672) = 84.89$, $MSE = 0.34$, $p < .001$, partial $\eta^2 = .28$ describing a very large effect; for flexible scoring, $F(3,672) = 79.30$, $MSE = 0.34$, $p < .001$, partial $\eta^2 = .26$ describing a very large effect. Nouns were more often translated correctly than the other parts of speech. This pattern is usual and is often attributed to the more concrete nature of many nouns.

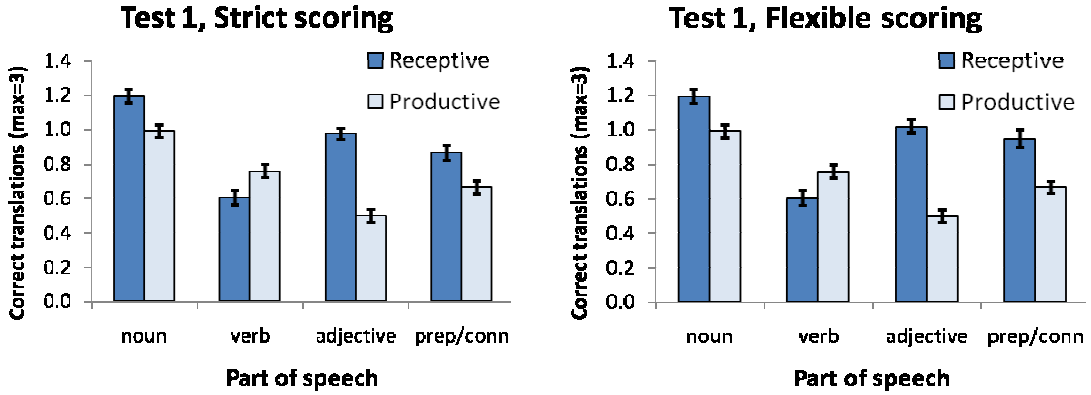


Error bars are ± 1 standard error of the mean.

Figure 2: Effect of part of speech

Part of speech was also involved in several interactions:

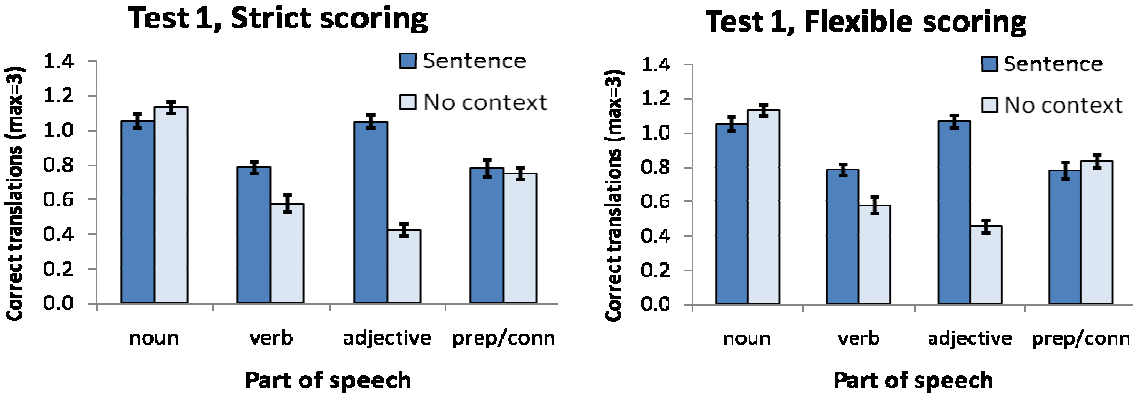
- with translation direction, for strict scoring, $F(3,672) = 45.88$, $MSE = 0.30$, $p < .001$, partial $\eta^2 = .17$ describing a large effect; for flexible scoring, $F(3,672) = 52.59$, $MSE = 0.31$, $p < .001$, partial $\eta^2 = .19$ describing a large effect. Verbs did not conform to the usual pattern wherein receptive translations are usually better than productive translations.



Error bars are ± 1 standard error of the mean.

Figure 3: Interactions between part of speech and translation direction

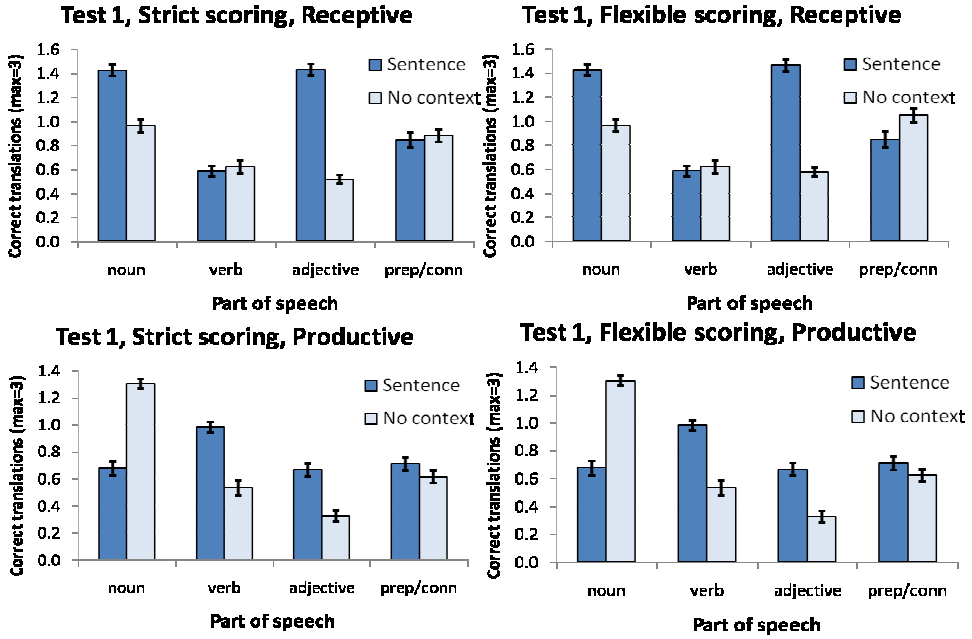
- with school, for strict scoring, $F(6,672) = 8.49$, $MSE = 0.34$, $p < .001$, partial $\eta^2 = .07$; for flexible scoring, $F(6,672) = 9.09$, $MSE = 0.34$, $p < .001$, partial $\eta^2 = .08$.
- with both translation direction and school, for strict scoring, $F(6,672) = 14.03$, $MSE = 0.30$, $p < .001$, partial $\eta^2 = .11$; for flexible scoring, $F(6,672) = 14.59$, $MSE = 0.31$, $p < .001$, partial $\eta^2 = .12$.
- with context, for strict scoring, $F(3,672) = 55.84$, $MSE = 0.36$, $p < .001$, partial $\eta^2 = .20$, describing a large effect; for flexible scoring, $F(3,672) = 60.90$, $MSE = 0.36$, $p < .001$, partial $\eta^2 = .21$, describing a large effect. Context provided a benefit for verbs and adjectives, but not for nouns and prepositions/connectives.



Error bars are ± 1 standard error of the mean.

Figure 4: Interactions between part of speech and sentence context

- with both context and school, for strict scoring, $F(6,672) = 12.06$, $MSE = 0.36$, $p < .001$, partial $\eta^2 = .10$; for flexible scoring, $F(6,672) = 12.05$, $MSE = 0.36$, $p < .001$, partial $\eta^2 = .10$.
- with both translation direction and context, for strict scoring, $F(3,672) = 80.94$, $MSE = 0.32$, $p < .001$, partial $\eta^2 = .27$, describing a very large effect; for flexible scoring, $F(3,672) = 88.73$, $MSE = 0.32$, $p < .001$, partial $\eta^2 = .28$, describing a very large effect.



Error bars are ± 1 standard error of the mean.

Figure 5: Interactions between part of speech, translation direction and sentence context

- and the four-way interaction with translation direction, context, and school, for strict scoring, $F(6,672) = 6.34$, $MSE = 0.32$, $p < .001$, partial $\eta^2 = .05$; for flexible scoring, $F(6,672) = 7.71$, $MSE = 0.32$, $p < .001$, partial $\eta^2 = .06$.

Phrases: Separate analyses were run for translation of phrases; the three-way ANOVA included school, gender, and translation direction.

On the first test, for strict scoring, there was a small but significant main effect of gender wherein girls produced more correct translations ($M = .39$, $SEM = .04$) than did boys ($M = .19$, $SEM = .05$, maximum possible score = 3), $F(1,224) = 9.75$, $MSE = .40$, $p = .002$, partial $\eta^2 = .04$. School and translation direction interacted significantly; pupils at School 2 were slightly better with productive translations than with receptive ones whereas the other two schools showed the usual pattern of better translation in the receptive direction.

Test 1, Strict scores, Phrase translation

School	Receptive	Productive
School 1	.33 *(.07)*	.19 *(.07)*
School 2	.24 *(.06)*	.45 *(.06)*
School 3	.34 *(.08)*	.19 *(.08)*
Overall	.30 *(.04)*	.28 *(.04)*

Maximum possible score was 3.

Table 14: Paper 1 results for phrase translations with strict scoring

For flexible scoring the pattern was similar, with a notable difference. Girls produced significantly more correct translations, but the difference was smaller with flexible scoring than that observed with strict scoring, $F(1,224) = 3.96$, $MSE = .32$, $p = .048$, partial $\eta^2 = .02$. For girls, the mean was .75 of a possible 3 phrases ($SEM = .04$); for boys the mean was .63 ($SEM = .04$).

With flexible scoring, many more receptive translations were scored as correct, whereas productive translations were essentially unchanged from the strict scoring. The unusual pattern for School 2, wherein students produced more correct translations in the productive direction, was replaced by the more customary advantage for receptive translations, albeit a smaller advantage than that observed in the other two schools. The main effect of translation direction was statistically significant, $F(1,224) = 409.32$, $MSE = .17$, $p < .001$, partial $\eta^2 = .65$ describing an extremely large effect. The interaction with school was also significant, $F(2,224) = 10.50$, $MSE = .17$, $p < .001$, partial $\eta^2 = .09$ describing a moderate effect.

Test 1, Flexible scores, Phrase translation

School	Receptive	Productive
School 1	1.06 *(.05)*	.19 *(.07)*
School 2	1.03 *(.04)*	.45 *(.06)*
School 3	1.20 *(.06)*	.19 *(.08)*
Overall	1.10 *(.03)*	.04 *(.04)*

Maximum possible score was 3.

Table 15: Paper 1 results for phrase translations with flexible scoring

Spelling: Spelling performance of correctly translated items on the first test is summarised in the following table. The maximum possible value for each cell would be 3 if all three items were correctly translated, but on the first test, at the beginning of the year, fewer than 10% of the pupils translated all three items correctly for any of the receptive translations and fewer than 4% translated all three items correctly for any of the productive translations.

The lower half of the table reports a conditionalised spelling score – the percentage of correctly translated items that were spelled correctly. The number of pupils included in each of these percentages is reported alongside the mean. Some pupils did not translate any of the three items correctly; for these it is not possible to calculate a percentage correctly spelled, so they are omitted from the data.

	Receptive – English to German						Productive – German to English					
Word type	No context			Sentence context			No context			Sentence context		
Spelling scores												
Nouns		0.9	*(0.79)*		1.5	*(0.69)*		1.1	*(0.38)*		0.2	*(0.49)*
Verbs		0.6	*(0.85)*		0.6	*(0.64)*		0.5	*(0.72)*		0.9	*(0.54)*
Adjectives		0.6	*(0.71)*		1.1	*(0.67)*		0.2	*(0.44)*		0.6	*(0.63)*
Prepositions and connectives		1.0	*(0.84)*		0.9	*(0.94)*		0.6	*(0.69)*		0.6	*(0.70)*
Phrases		1.1	*(0.46)*					0.2	*(0.44)*			
Spelling %	*N*			*N*			*N*			*N*		
Nouns	167	94.1	*(22.34)*	221	99.7	*(3.16)*	227	93.3	*(17.76)*	116	30.0	*(42.69)*
Verbs	105	99.0	*(9.76)*	117	99.1	*(9.25)*	91	88.6	*(25.92)*	199	87.7	*(30.21)*
Adjectives	123	96.3	*(18.30)*	208	75.4	*(31.64)*	59	61.9	*(40.83)*	133	83.7	*(34.78)*
Prepositions and connectives	166	88.4	*(28.81)*	129	97.5	*(13.68)*	127	93.7	*(23.56)*	120	84.6	*(32.28)*
Phrases	224	99.3	*(12.05)*				59	59.3	*(46.86)*			

Table 16: Spelling scores for Paper 1 according to part of speech, translation direction and sentence context

4.5 Performance on the second test

The strict and flexible translation scores from the second test are summarised here. Means are reported (with standard deviations in parentheses). The maximum possible value for each cell is 3.

	Receptive - English to German				Productive - German to English			
Word type	No context		Sentence context		No context		Sentence context	
Strict scores								
Nouns	2.1	*(0.72)*	2.2	*(0.67)*	2.0	*(0.78)*	2.2	*(0.70)*
Verbs	1.7	*(0.88)*	1.8	*(0.90)*	1.5	*(0.97)*	1.9	*(0.81)*
Adjectives	1.4	*(0.73)*	2.1	*(0.62)*	1.4	*(0.83)*	1.4	*(0.89)*
Prepositions and connectives	1.7	*(0.90)*	2.2	*(0.80)*	1.5	*(0.81)*	2.2	*(0.90)*
Phrases	0.9	*(0.88)*			1.0	*(1.03)*		
Flexible scores								
Nouns	2.1	*(0.72)*	2.2	*(0.67)*	2.0	*(0.78)*	2.2	*(0.70)*
Verbs	1.7	*(0.88)*	1.8	*(0.86)*	1.5	*(0.97)*	1.9	*(0.81)*
Adjectives	1.7	*(0.88)*	2.2	*(0.67)*	1.4	*(0.83)*	1.4	*(0.89)*
Prepositions and connectives	2.2	*(0.84)*	2.2	*(0.80)*	1.5	*(0.81)*	2.2	*(0.90)*
Phrases	1.7	*(0.77)*			1.0	*(1.03)*		

Table 17: Paper 2 scores for different parts of speech, translation direction and sentence context

In order to detect interactions as well as main effects, school, gender, translation direction (receptive or productive), sentence context, and part of speech (noun, verb, adjective, preposition or connective) were investigated with a five-way mixed ANOVA for pupils' translations at the end of the year. Translation of phrases was not included in this analysis. Note that this analysis is concerned with pupils' ability to translate the words at the end of the year, not with how much they have improved

during the year. Differences at the beginning of the year are likely to influence these data in a variety of ways.

School: School had a significant effect on pupils' initial performance, for strict scoring $F(2,224) = 30.27$, $MSE = 3.32$, $p < .001$, partial $\eta^2 = .21$ describing a very large effect; for flexible scoring $F(2,224) = 31.15$, $MSE = 3.35$, $p < .001$, partial $\eta^2 = .22$ describing a very large effect. The relative positions of School 3 and School 2 reversed during the year. Tukey posthoc comparisons show that the students at School 2 produced more correct translations than did the students at School 3, $p = .001$. The latter produced more correct translations than did students at School 1, $p = .001$. School 2's pupils also produced significantly more translations than School 1's, $p < .001$. Later in the report, it is noted that school is also related to phase-6 usage.

	Strict		Flexible	
School	Mean	SEM	Mean	SEM
School 1	1.52	0.055	1.57	0.055
School 2	2.08	0.047	2.14	0.047
School 3	1.78	0.061	1.85	0.061

The maximum possible score was 3 for each cell.
Table 18: Paper 2 scores for each school

Gender: The effect of gender was not significant at the end of the year. For strict scoring, $F(1,224) = 2.12$, $MSE = 3.32$, $p = .147$, partial $\eta^2 = .009$; for flexible scoring, $F(1,224) = 2.59$, $MSE = 3.35$, $p = .109$, partial $\eta^2 = .011$.

	Strict		Flexible	
Gender	Mean	SEM	Mean	SEM
Girls	1.84	0.042	1.90	0.042
Boys	1.75	0.047	1.80	0.048

The maximum score was 3 for each cell.
Table 19: Paper 2 scores by gender

Translation direction: As is usual when learning foreign vocabulary, pupils were significantly better at receptive translations than productive ones; for strict scoring $F(1,224) = 51.61$, $MSE = 0.38$, $p < .001$, partial $\eta^2 = .19$ describing a large effect, for flexible scoring $F(1,224) = 151.42$, $MSE = 0.36$, $p < .001$, partial $\eta^2 = .40$ describing a very large effect.

	Strict		Flexible	
Direction	Mean	SEM	Mean	SEM
Receptive	1.87	0.032	1.98	0.032
Productive	1.72	0.034	1.72	0.034

The maximum possible score was 3 for each cell.
Table 20: Paper 2 scores by translation direction

The direction of the translation interacted with school. For strict scoring $F(2,224) = 16.15$, $MSE = 0.38$, $p < .001$, partial $\eta^2 = .13$ suggesting a large effect; for flexible scoring $F(2,224) = 17.91$, $MSE = 0.36$, $p < .001$, partial $\eta^2 = .14$ suggesting a large effect.

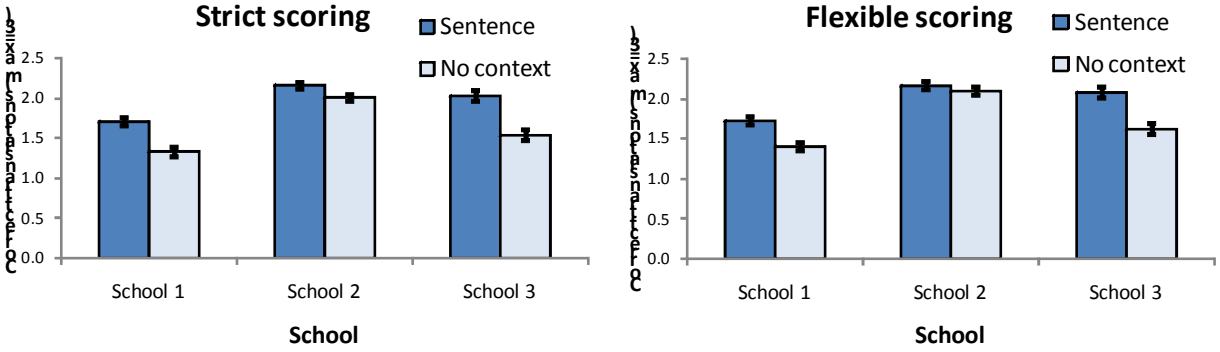
The direction of translation did not interact significantly with gender, but the three way interaction with school and gender was statistically significant; the effect was small and perhaps not worthy of interpretation. For strict scoring $F(2,224) = 5.10$, $MSE = 0.38$, $p = .007$, partial $\eta^2 = .04$; for flexible scoring $F(2,224) = 4.05$, $MSE = 0.36$, $p = .019$, partial $\eta^2 = .04$.

Context: Students translated items correctly significantly more often when a sentence was present for context than when it was not; for strict scoring, $F(1,224) = 272.53$, $MSE = 0.36$, $p < .001$, partial $\eta^2 = .55$ describing a very large effect; for flexible scoring, $F(1,224) = 181.15$, $MSE = 0.36$, $p < .001$, partial $\eta^2 = .45$ describing a very large effect. This effect might be due to the support provided by the context or to having more familiar or easier words offered with a sentence for context; further research could vary the test among the students so that the same words were offered with sentences for some students and without them for others.

	Strict		Flexible	
Context	Mean	SEM	Mean	SEM
Sentence	1.97	0.033	2.00	0.033
None	1.62	0.034	1.71	0.034

The maximum possible score was 3 for each cell.
Table 21: Paper 2 scores by sentence context

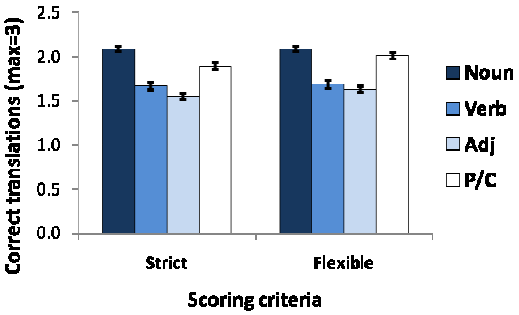
Context interacted significantly with school and for strict scoring with gender, but the three-way interaction was not significant. For the context x school interaction for strict scoring, $F(2,224) = 25.44$, $MSE = 0.36$, $p < .001$, partial $\eta^2 = .19$; for flexible scoring, $F(2,224) = 3.50$, $MSE = 0.56$, $p = .032$, partial $\eta^2 = .03$. The figure below illustrates the nature of the interaction, which is similar to that observed in the initial test: For pupils at School 2 the sentence context provided far less benefit.



Error bars are ± 1 standard error of the mean.

Figure 6: Interactions between school and sentence context

Part of speech: The part of speech played a significant role in initial test translations, for strict scoring, $F(3,672) = 98.91$, $MSE = 0.48$, $p < .001$, partial $\eta^2 = .31$ describing a very large effect; for flexible scoring, $F(3,672) = 19.99$, $MSE = 0.77$, $p < .001$, partial $\eta^2 = .08$ describing a moderate effect. Nouns were more often translated correctly than the other parts of speech. This pattern is usual and is often attributed to the more concrete nature of many nouns. In a departure from the pattern observed in paper 1, prepositions and connectives (P/C) were translated correctly almost as often as were nouns.

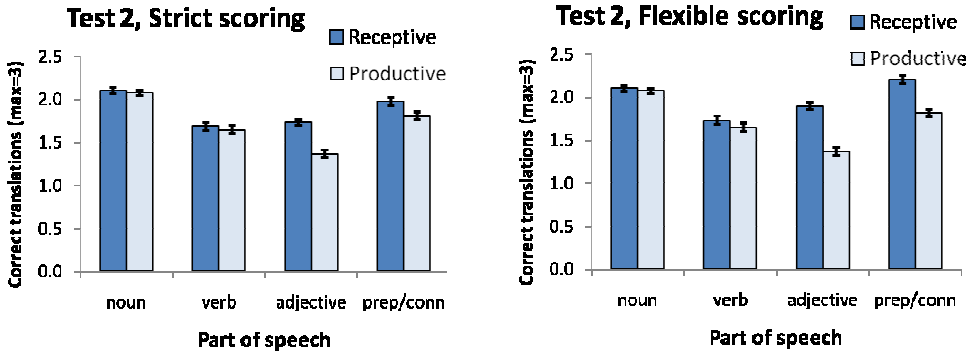


Error bars are ± 1 standard error of the mean.

Figure 7: Effect of part of speech

As for the first test, part of speech was also involved in several interactions:

- with translation direction, for strict scoring, $F(3,672) = 14.54$, $MSE = 0.35$, $p < .001$, partial $\eta^2 = .06$ describing a small to moderate effect, for flexible scoring, $F(3,672) = 10.74$, $MSE = 0.57$, $p < .001$, partial $\eta^2 = .05$ describing a small effect. On this test, neither nouns nor verbs conformed to the usual pattern where receptive translations are usually better than productive translations; on the first test nouns showed the usual difference but verbs did not.



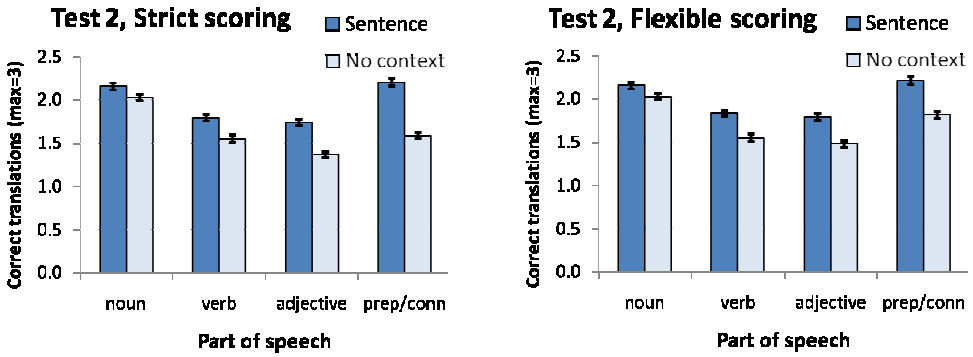
Error bars are ± 1 standard error of the mean.

Figure 8: Interactions between part of speech and translation direction

- with school, for strict scoring, $F(6,672) = 5.33$, $MSE = 0.48$, $p < .001$, partial $\eta^2 = .05$; for flexible scoring, $F(6,672) = 12.21$, $MSE = 0.77$, $p < .001$, partial $\eta^2 = .10$.

- with both translation direction and school, for strict scoring, $F(6,672) = 3.60$, $MSE = 0.35$, $p = .002$, partial $\eta^2 = .03$; for flexible scoring, $F(6,672) = 9.01$, $MSE = 0.57$, $p < .001$, partial $\eta^2 = .07$.

- with context, for strict scoring, $F(3,672) = 21.41$, $MSE = 0.43$, $p < .001$, partial $\eta^2 = .09$, describing a moderate effect; for flexible scoring, $F(3,672) = 27.72$, $MSE = 0.76$, $p < .001$, partial $\eta^2 = .11$, describing a moderate effect. Context provided little benefit for nouns, but improved translations for the other parts of speech. In test 1, context did not appear to help prepositions and connectives, but here there was a clear effect. The effect was less marked for adjectives in test 2 than it had been in test 1.

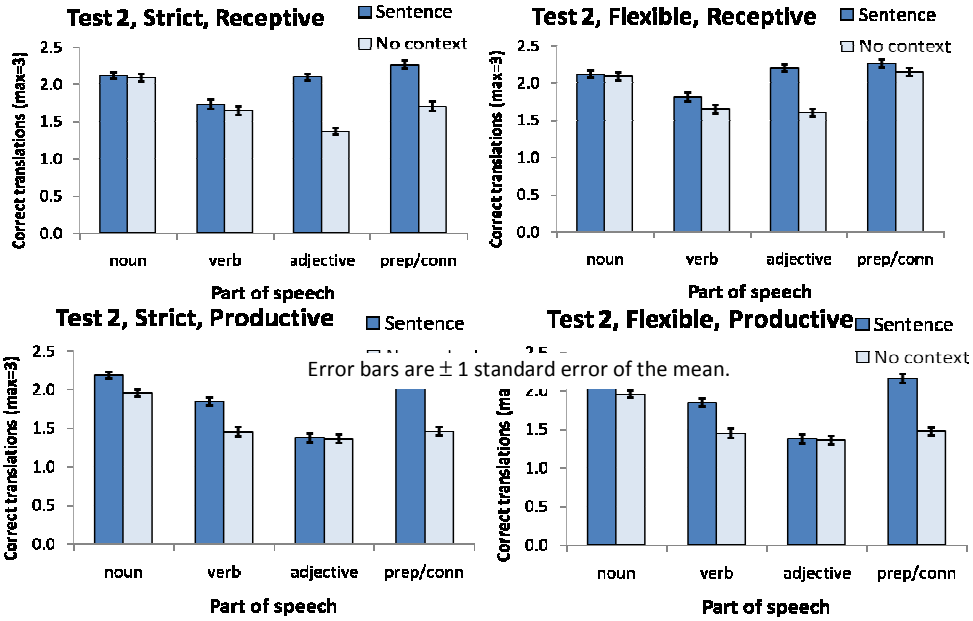


Error bars are ± 1 standard error of the mean.

Figure 9: Interactions between part of speech and sentence context

- with both context and school, for strict scoring, $F(6,672) = 4.79$, $MSE = 0.43$, $p < .001$, partial $\eta^2 = .04$, a small effect; for flexible scoring, $F(6,672) = 2.60$, $MSE = 0.76$, $p = .02$, partial $\eta^2 = .02$, a very small effect.

- with both translation direction and context, for strict scoring, $F(3,672) = 32.28$, $MSE = 0.36$, $p < .001$, partial $\eta^2 = .1327$, describing a large effect; for flexible scoring, $F(3,672) = 39.86$, $MSE = 0.60$, $p < .001$, partial $\eta^2 = .15$, describing a large effect.



Error bars are ± 1 standard error of the mean.

Figure 10: Interactions between parts of speech, sentence context and translation direction

- and the four-way interaction with translation direction, context, and school, for strict scoring, $F(6,672) = 3.91$ $MSE = 0.36$, $p = .001$, partial $\eta^2 = .03$, describing a very small effect; for flexible scoring, $F(6,672) = 2.67$, $MSE = 0.32$, $p = .046$, partial $\eta^2 = .01$, describing a very small effect.

Phrases: Separate analyses were run for translation of phrases; the three-way ANOVA included school, gender, and translation direction.

On the second test, for strict scoring, the effect of gender observed in test 1 had been lost; girls produced roughly the same number of correct translations as boys with both scoring metrics.

Test 2, Phrase translation

	Strict		Flexible	
Gender	Mean	SEM	Mean	SEM
Girls	0.90	0.07	1.30	0.06
Boys	0.94	0.08	1.33	0.07

The maximum possible score was 3 for each cell.

Table 22: Paper 2 results of phrase translation by gender

For strict scoring there were significantly more translations in the productive (German-English) direction than in the receptive (English-German) direction, $F(1,224) = 5.53$, $MSE = .52$, $p = .02$, partial $\eta^2 = .02$. This unusual pattern can be examined more closely in the significant interaction, $F(2,224) = 3.21$, $MSE = .52$, $p = .042$, partial $\eta^2 = .03$. Two of the three schools showed this pattern. The main effect of school was also statistically significant, $F(2,224) = 22.84$, $MSE = 1.09$, $p < .001$, partial $\eta^2 = .17$.

Test 2, Strict scores, Phrase translation

School	Receptive		Productive	
School 1	.71	*(.10)*	.62	*(.12)*
School 2	1.24	*(.09)*	1.50	*(.10)*
School 3	.55	*(.11)*	.88	*(.13)*
Overall	.83	*(.06)*	1.00	*(.07)*

Maximum possible score was 3.

Table 23: Paper 2 results of phrase translation by school

For flexible scoring, there were significantly more translations in the receptive (English-German) direction ($M = 1.63$, $SEM = .05$) than in the productive (German-English) direction ($M = 1.00$, $SEM = .07$), $F(1,224) = 88.84$, $MSE = .47$, $p < .001$, partial $\eta^2 = .28$. The effect was quite large and the pattern is quite different from that observed with strict scoring. The main effect of school was also statistically significant, $F(2,224) = 24.17$, $MSE = 0.97$, $p < .001$, partial $\eta^2 = .18$. Pupils at School 2 produced the greatest number of correct translations ($M = 1.74$, $SEM = .07$) in contrast to School 3 ($M = 1.19$, $SEM = .09$) and School 1 ($M = 1.01$, $SEM = .08$).

Spelling: Spelling scores for the second test are summarised here. Conditionalised scores in the lower half of the table show that at the end of the year, after study and phase-6 activity, correct translations were correctly spelled the vast majority of the time.

	Receptive – English to German						Productive – German to English					
Word type	No context			Sentence context			No context			Sentence context		
Spelling scores												
Nouns		2.1	*(0.75)*		2.1	*(0.67)*		1.8	*(0.81)*		1.5	*(0.94)*
Verbs		1.7	*(0.88)*		1.8	*(0.87)*		1.5	*(0.97)*		1.8	*(0.82)*
Adjectives		1.7	*(0.88)*		1.5	*(0.77)*		1.2	*(0.75)*		1.4	*(0.87)*
Prepositions and connectives		1.9	*(0.88)*		2.2	*(0.82)*		1.5	*(0.77)*		2.0	*(0.91)*
Phrases		1.6	*(0.76)*					0.9	*(1.02)*			
Spelling %	*N*			*N*			*N*			*N*		
Nouns	226	97.6	*(13.25)*	226	99.7	*(4.43)*	230	87.6	*(21.47)*	226	67.8	*(34.46)*
Verbs	221	99.8	*(3.36)*	218	96.9	*(13.70)*	194	96.9	*(12.77)*	222	94.4	*(17.21)*
Adjectives	217	99.1	*(6.74)*	227	70.4	*(29.82)*	195	86.3	*(24.89)*	200	95.2	*(17.49)*
Prepositions and connectives	222	90.3	*(20.73)*	224	96.7	*(12.53)*	216	97.6	*(10.82)*	218	92.8	*(17.14)*
Phrases	226	97.8	*(9.94)*				139	80.9	*(35.64)*			

Table 24: Spelling scores in Paper 2 by parts of speech, translation direction and sentence context

4.6 Improvement: Test 2 minus Test 1

Improvements on Test 2 over Test 1 reflect the use of phase-6 plus all other learning activities that took place between the two tests (and this is likely to include uses of boxes and cards, uses of word lists in books, and checking translations with parents and peers). In addition some of the change will be due to the error inherent in each measure: students may fail to translate something that they actually know and they may successfully translate something more or less by chance.

The improvements based on strict and flexible translation scores from the second test are summarised here. Means are reported (with standard deviations in parentheses). The maximum possible value for each cell is 3, where no items were initially correct and all were correct on the final test; the minimum possible value for each cell is -3, where all items were originally correct and none were finally correct.

	Receptive - English to German				Productive - German to English			
Word type	No context		Sentence context		No context		Sentence context	
Strict scores								
Nouns	1.1	*(1.07)*	0.7	*(0.88)*	0.8	*(1.01)*	1.5	*(1.03)*
Verbs	1.1	*(1.04)*	1.2	*(1.00)*	0.9	*(1.04)*	0.9	*(0.84)*
Adjectives	0.8	*(0.81)*	0.7	*(0.86)*	1.1	*(1.05)*	0.7	*(1.00)*
Prepositions and connectives	0.8	*(1.00)*	1.4	*(1.15)*	0.8	*(0.92)*	1.5	*(1.01)*
Phrases	0.6	*(0.99)*			0.7	*(1.12)*		
Flexible scores								
Nouns	1.1	*(1.07)*	0.7	*(0.88)*	0.8	*(1.01)*	1.5	*(1.03)*
Verbs	1.1	*(1.04)*	1.3	*(0.98)*	0.9	*(1.04)*	0.9	*(0.84)*
Adjectives	1.0	*(0.84)*	0.7	*(0.84)*	1.1	*(1.05)*	0.7	*(1.00)*
Prepositions and connectives	1.1	*(1.04)*	1.4	*(1.15)*	0.9	*(0.90)*	1.5	*(1.01)*
Phrases	0.6	*(0.83)*			0.7	*(1.12)*		

Table 25: Differences between test scores by part of speech, transalation direction and sentence context

Students generally improved in all cells, as evidenced by the positive mean values above and by the histograms below.

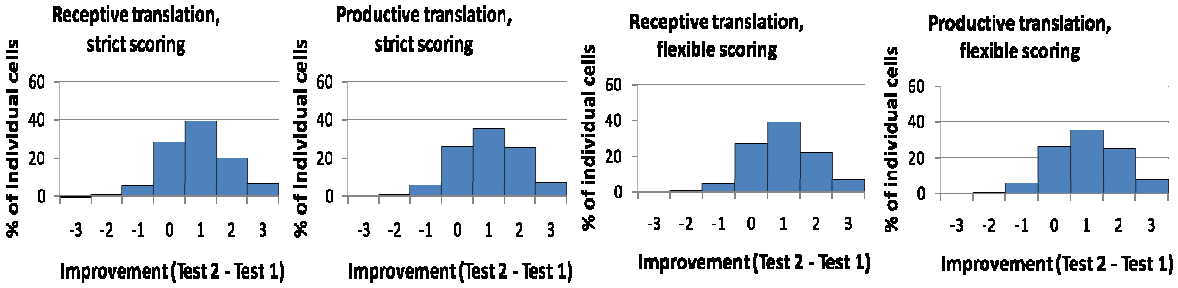


Figure 11: Improvements shown according to translation direction

Error bars are ± 1 standard error of the mean.

In order to detect interactions as well as main effects, the influence of translation direction (receptive or productive), sentence context, and part of speech (noun, verb, adjective, preposition/connective) were investigated with a five-way mixed ANOVA for the change in pupils' translations between the first and second tests. Translation of phrases was not included in this analysis, but was analysed separately.

School: Schools differed significantly in how much their students improved, for strict scoring $F(2,224) = 17.70$, $MSE = 4.14$, $p < .001$, partial $\eta^2 = .14$, for flexible scoring $F(2,224) = 17.25$, $MSE = 4.19$, $p < .001$, partial $\eta^2 = .13$. Pupils at School 2 improved substantially more than those at the other two schools; the advantage was significant according to a Tukey test, $p < .001$ for both strict and flexible scores.

	Strict		Flexible	
School	Mean	SEM	Mean	SEM
School 3	0.82	0.068	0.87	0.069
School 2	1.26	0.052	1.29	0.053
School 1	0.86	0.061	0.90	0.062

The range of possible scores was -3 to 3 for each cell.

Table 26: Improvements shown by school

Gender: Boys improved slightly but significantly more than did girls, for strict scoring $F(1,224) = 6.36$, $MSE = 4.14$, $p = .012$, partial $\eta^2 = .03$, for flexible scoring $F(1,224) = 6.13$, $MSE = 4.19$, $p = .014$, partial $\eta^2 = .03$.

	Strict		Flexible	
Gender	Mean	SEM	Mean	SEM
Boys	1.07	0.053	1.11	0.053
Girls	0.89	0.046	0.93	0.047

The range of possible scores was -3 to 3.

Table 27: Improvements shown by gender

Translation direction: Overall improvements for receptive and productive translations were very similar, at about 1 word out of a maximum possible 3. For strict scoring average improvements were 0.97 ($SEM = 0.038$) and 0.99 ($SEM = 0.038$) for receptive and productive translations, respectively. For flexible scoring average improvements were 1.04 ($SEM = 0.038$) and 1.00 ($SEM = 0.038$).

The pattern was not the same across schools, though, and the interaction was significant, for strict scoring $F(2,224) = 17.83$, $MSE = 0.59$, $p < .001$, partial $\eta^2 = .14$, for flexible scoring $F(2,224) = 18.42$, $MSE = 0.60$, $p < .001$, partial $\eta^2 = .14$. For both School 1 and School 2 improvements were greater for receptive translations whereas for School 2 improvements were greater for productive translations.

Translation direction did not interact significantly with gender, but the three way interaction between school, gender and translation direction was statistically significant, for strict scoring $F(2,224) = 7.56$, $MSE = 0.59$, $p = .001$, partial $\eta^2 = .06$, for flexible scoring $F(2,224) = 6.89$, $MSE = 0.60$, $p = .001$, partial $\eta^2 = .06$.

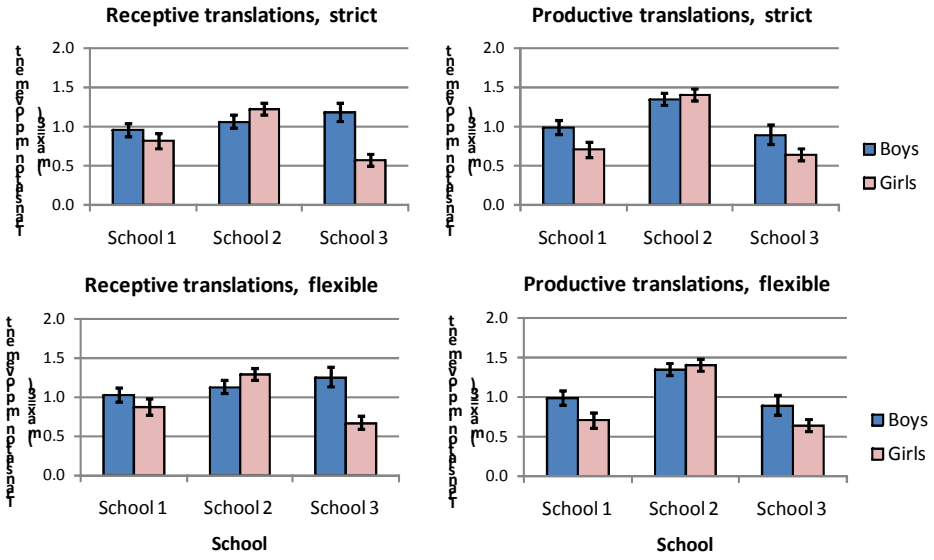


Figure 12: Interactions between school, translation direction and gender

Context: Isolated translations improved slightly and significantly more than translations presented in the context of a sentence, for strict scoring $F(1,224) = 30.56$, $MSE = 0.58$, $p < .001$, partial $\eta^2 = .12$, for flexible scoring $F(2,224) = 17.37$, $MSE = 0.56$, $p < .001$, partial $\eta^2 = .07$. For strict scoring, pupils improved by an average of 0.91 words ($SEM = 0.038$) when a sentence was provided for context; they improved by 1.05 words ($SEM = 0.037$) when no context was provided. For flexible scoring, translations improved by 0.96 ($SEM = 0.038$) with context and by 1.07 ($SEM = 0.037$) without.

For flexible scoring, context interacted significantly with school, $F(2,224) = 3.50$, $MSE = 0.56$, $p = .032$, partial $\eta^2 = .03$ describing a small effect.

Context interacted significantly with translation direction, for strict scoring $F(1,224) = 22.04$, $MSE = 0.55$, $p < .001$, partial $\eta^2 = .09$, for flexible scoring $F(1,224) = 38.04$, $MSE = 0.55$, $p < .001$, partial $\eta^2 = .15$. Context contributed more to improvement in receptive translations.

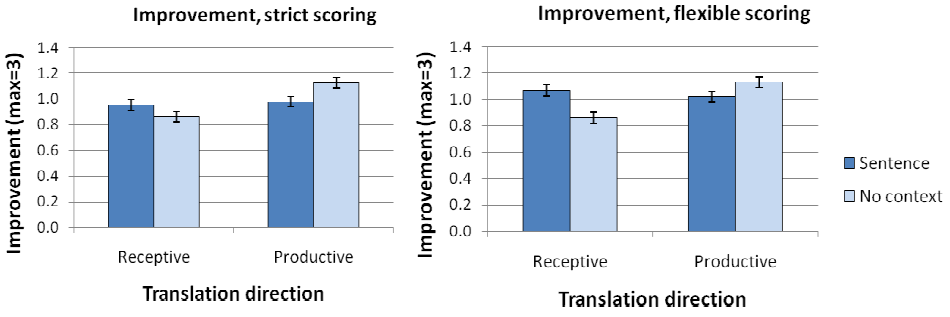


Figure 13: Interactions between translation direction and sentence context

This interaction with context was different across different schools, for strict scoring $F(2,224) = 7.27$, $MSE = 0.55$, $p = .001$, partial $\eta^2 = .06$, for flexible scoring $F(2,224) = 7.15$, $MSE = 0.55$, $p < .001$, partial $\eta^2 = .06$. For students at School 1, providing context was not associated with greater improvement. Students at School 2 showed greater improvement with context for receptive but not productive translations. For students at School 3, the presence or absence of context was not related to improvement for receptive translations, but more improvement was shown for words without context translated from German to English – the more difficult, productive translation. This pattern is illustrated below for flexible scoring – the same pattern was observed with strict scoring.

Figure 14: Interactions between school, translation direction and sentence context

Part of speech: Improvements varied significantly by part of speech, for strict scoring $F(3,224) = 18.43$, $MSE = 0.76$, $p < .001$, partial $\eta^2 = .08$, for flexible scoring $F(3,224) = 19.99$, $MSE = 0.77$, $p < .001$, partial $\eta^2 = .10$. The greatest improvements were observed for prepositions and connectives.

The effects of part of speech interacted with other factors in terms of improvements:

School x part of speech interaction – a moderate sized effect ($\eta^2 = .09$ and .10 for strict and lenient scoring respectively) – The pattern of improvements was similar at School 1 and School 3: nouns, verbs and adjectives all improved roughly the same amount and prepositions and connectives improved noticeably more. The pattern at School 2 was quite different as shown in the graph. The patterns for strict scoring were quite similar, showing slightly lower improvements for most cells. For strict scoring $F(6,224) = 11.11$, $MSE = 0.76$, $p < .001$, partial $\eta^2 = .09$, for flexible scoring $F(6,224) = 12.21$, $MSE = 0.77$, $p < .001$, partial $\eta^2 = .10$.

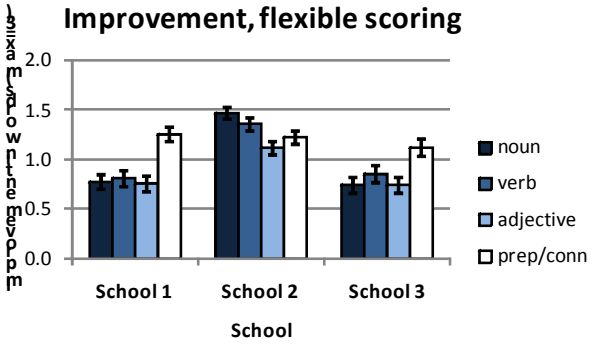


Figure 15: Effect of parts of speech and school

Translation direction x part of speech interaction - a small sized effect ($\eta^2 = .04$ and .05 for strict and lenient scoring respectively) – For receptive translations the improvements were greater for verbs than nouns; for productive translations the pattern was reversed. The graph shows the pattern for flexible scoring; the same pattern appeared for strict scoring. For strict scoring $F(3,224) = 9.49$, $MSE = 0.57$, $p < .001$, partial $\eta^2 = .04$, for flexible scoring $F(3,224) = 10.74$, $MSE = 0.57$, $p < .001$, partial $\eta^2 = .05$.

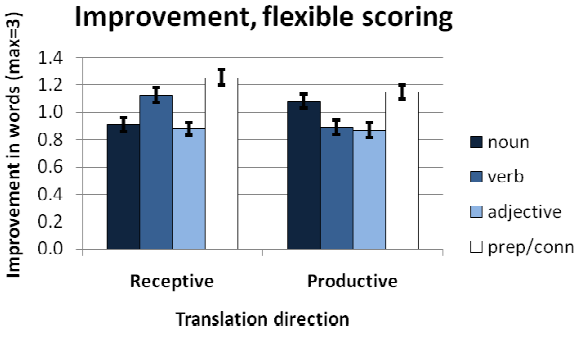
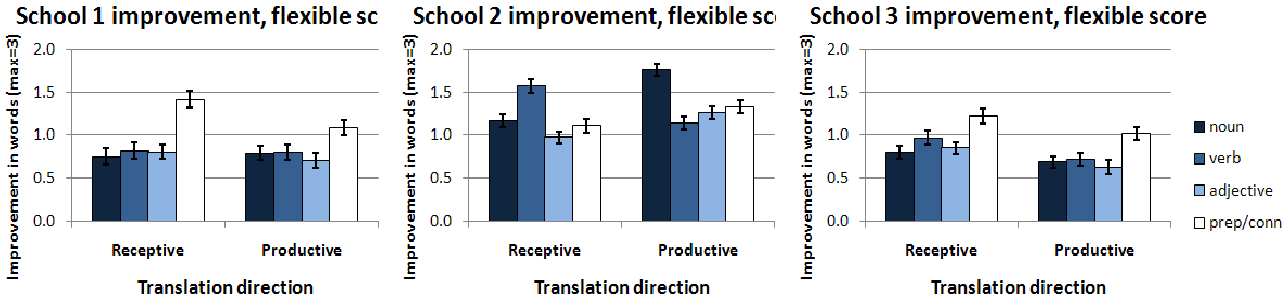


Figure 16: Effect of parts of speech and translation direction

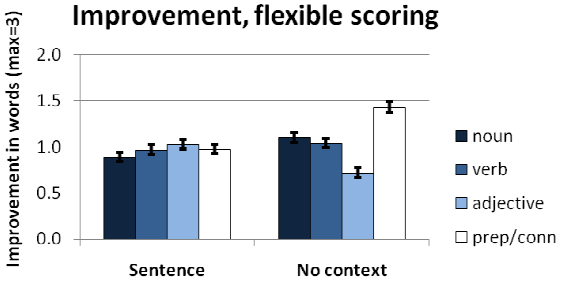
School and translation direction x part of speech interaction - a moderate to small effect (η^2 = .08 and .06 for strict and lenient scoring respectively) – The relationship between translation direction and part of speech was different for different schools. School 1 and School 3 showed similar patterns of improvement, but students at School 2 showed the greatest improvements for verbs in receptive translations and for nouns in productive ones. For strict scoring $F(6,224) = 9.49$, $MSE = 0.57$, $p < .001$, partial $\eta^2 = .08$, for flexible scoring $F(6,224) = 9.01$, $MSE = 0.57$, $p < .001$, partial $\eta^2 = .07$.



Error bars are ± 1 standard error of the mean.

Figure 17: Effect of school, parts of speech and translation direction

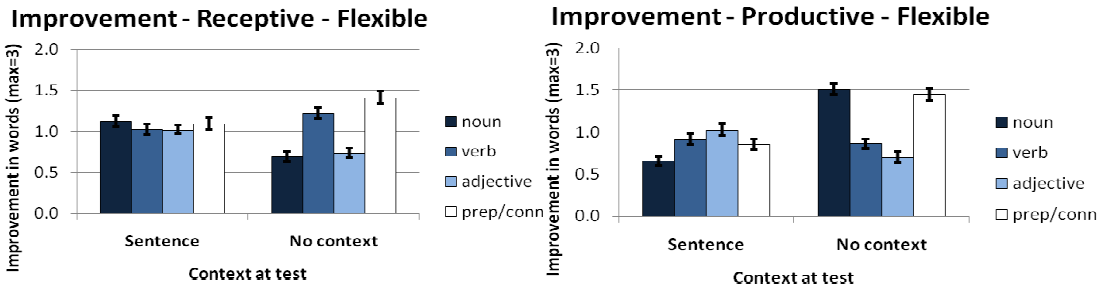
Context x part of speech interaction - a moderate to large effect (η^2 = .13 and .11 for strict and lenient scoring respectively) – When tested within a sentence, all parts of speech showed similar overall levels of improvement whereas when tested in isolation, adjectives improved less and prepositions and connectives improved more. The pattern for flexible scoring is shown here; for strict scoring the pattern was the same. For strict scoring $F(3,224) = 33.97$, $MSE = 0.77$, $p < .001$, partial $\eta^2 = .13$, for flexible scoring $F(3,224) = 27.72$, $MSE = 0.76$, $p < .001$, partial $\eta^2 = .11$.



Error bars are ± 1 standard error of the mean.

Figure 18: Effect of parts of speech and sentence context

School and context x part of speech interaction – a very small effect (η^2 = .03 and .02 for strict and lenient scoring respectively) – The above pattern varied slightly by school. For strict scoring $F(6,224) = 2.82$, $MSE = 0.77$, $p = .01$, partial $\eta^2 = .03$, for flexible scoring $F(6,224) = 2.60$, $MSE = 0.76$, $p = .02$, partial $\eta^2 = .02$.



Error bars are ± 1 standard error of the mean.

Figure 19: Effect of translation direction, parts of speech and sentence context

Translation direction and context x part of speech interaction - a large effect (η^2 = .15 for both strict and flexible scoring) – The above pattern with respect to context and part of speech averaged over a difference related to the direction of the translation. As can be seen in the graphs below, improvements for a given part of speech (for example, nouns) varied based on both the context and the translation direction. For example, for nouns testing with a sentence for context led to

greater improvements when testing English to German and to less improvement when testing German to English.

Translation direction, context and gender x part of speech interaction - a very small effect (η^2 = .02 and .01 for strict and lenient scoring respectively) – The above pattern varied slightly by gender.

Phrases: A separate three way ANOVA (school x gender x translation direction) was run to analyse improvement on translating the phrases.

Students improved slightly but significantly more for productive translations than for receptive ones; for strict scoring $F(1,224) = 6.29$, $MSE = 0.59$, $p = .013$, partial $\eta^2 = .03$, for flexible scoring $F(1,224) = 7.04$, $MSE = 0.53$, $p = .009$, partial $\eta^2 = .03$.

	Strict				Flexible			
School	Receptive		Productive		Receptive		Productive	
School 1	0.38	*(0.11)*	0.43	*(0.13)*	0.34	*(0.09)*	0.43	*(0.13)*
School 2	1.00	*(0.09)*	1.05	*(0.11)*	0.95	*(0.08)*	1.05	*(0.11)*
School 3	0.21	*(0.12)*	0.69	*(0.15)*	0.30	*(0.10)*	0.69	*(0.15)*
Overall	0.53	*(0.06)*	0.72	*(0.08)*	0.53	*(0.05)*	0.72	*(0.08)*

Mean and *(SD)* for each cell; possible values are -3 to 3.

Table 28: Improvement scores for phrase translation according to school and translation direction

The main effect of school was also significant for both strict and flexible scoring; for strict scoring $F(2,224) = 13.54$, $MSE = 1.41$, $p < .001$, partial $\eta^2 = .11$, for flexible scoring $F(2,224) = 14.06$, $MSE = 1.24$, $p < .001$, partial $\eta^2 = .11$.

For strict scoring, but not for flexible scoring, translation direction interacted with school; for strict scoring $F(2,224) = 3.18$, $MSE = 0.59$, $p = .043$, partial $\eta^2 = .03$, for flexible scoring $F(2,224) = 1.68$, $MSE = 0.53$, $p = .19$, partial $\eta^2 = .02$. For School 3 the improvement was markedly greater for productive translations than for receptive ones; for the other schools the difference was in the same direction, but was quite small.

4.7 Usage levels of phase-6 related to improvements at school level

Pupils who reported that they had used phase-6 did not perform significantly differently from pupils who reported that they had not used phase-6; this applied to Test 1 performance (that is, how well they did initially did not relate to whether or not they used phase-6), to Test 2 performance, and to the improvement scores. Comparing pupils who used phase-6 about daily with those who did not gained the same result.

Based on large differences between schools, it would appear that other aspects of English language learning may well play a much larger role than has their use of phase-6. Following this line of thought, comparisons can be run one school at a time.

For pupils at School 3 there is a suggestion that phase-6 has made a difference, especially when the test item did not provide a sentence for context; for the other schools there was no such suggestion. However, the suggestion was weak as there were only 9 pupils who reported using phase-6 daily and 56 who did not. The pattern is similar when comparing reported use (32 pupils) and non-use (33 pupils). Most observed differences - and all significant differences - are in the direction predicted: phase-6 users produce more correct translations on Test 2 than did the pupils who did not use phase-6. The bulk of the significant differences were observed in the Test 2 scores - not in the improvement scores. A few significant differences were observed in the Test 1 scores. So the picture is somewhat unclear – there is a very weak indication that the better pupils were the ones who decided to use phase-6 and that their final test performance stretched their lead out a little more - enough to make more significant differences. But their improvement scores were not significantly better than the

improvement scores of the others. Of course, improvement scores are a somewhat vexed measure - a pupil who did poorly on Test 1 (for example, who got 0 correct words) could have a much higher improvement score (up to 3) than one who did well on Test 1 (for example, if 2 were correct on Test 1, the improvement score could not be more than 1). To address this issue, it will be worth considering a % improvement score in subsequent analyses.

A summary of the Test 2 differences observed for School 3 is shown in the table following. This includes the measure, the mean (and standard deviation) for the 56 non-daily users, the mean (and standard deviation) for the 9 daily users, and the result of a t test to identify significant differences. The significance (*p*) values reported are two-tailed – the significance level of the one-tailed test is half of these values. However, given the large number of t-tests, the criterion of .05 should probably be adjusted downwards (although a Bonferroni adjustment is not necessarily highly useful within this context). There were 63 degrees of freedom for each t test. In all cases where a significant difference was observed, the students who used phase 6 daily recalled more words than did the other students. These are flagged with 'phase-6' in the following table.

	Test 2, strict scoring		Test 2, flexible scoring		Test 2, spelling	
	Receptive	Productive	Receptive	Productive	Receptive	Productive
Noun, no context						
non-daily	2.2 (0.63)	1.8 (0.64)	2.2 (0.63)	1.8 (0.64)	2.2 (0.64)	1.4 (0.66)
daily	2.2 (0.83)	1.9 (0.78)	2.2 (0.83)	1.9 (0.78)	2.1 (1.05)	1.3 (1.00)
t, p	0.04, 0.97	0.59, 0.56	0.04, 0.97	0.59, 0.56	0.33, 0.74	0.44, 0.66
Verb, no context						
non-daily	1.6 (0.78)	1.3 (0.68)	1.6 (0.78)	1.3 (0.68)	1.6 (0.78)	1.2 (0.65)
daily	2.1 (0.93)	1.2 (0.67)	2.1 (0.93)	1.2 (0.67)	2.1 (0.93)	1.2 (0.67)
t, p	1.87, 0.07	0.26, 0.80	1.87, 0.07	0.26, 0.80	1.93, 0.06	0.03, 0.97
Adjective, no context						
non-daily	1.3 (0.61)	1.3 (0.80)	1.4 (0.68)	1.3 (0.80)	1.4 (0.68)	1.1 (0.72)
daily	1.7 (0.71)	1.8 (0.97)	1.8 (0.67)	1.8 (0.97)	1.7 (0.71)	1.4 (0.88)
t, p	1.55, 0.13	1.73, 0.09	1.66, 0.10	1.73, 0.09	1.19, 0.24	1.13, 0.26
Preposition and connective, no context						
non-daily	1.7 (0.90)	1.2 (0.64)	2.2 (0.81)	1.3 (0.67)	2.1 (0.83)	1.2 (0.65)
daily	1.3 (0.71)	1.4 (0.73)	2.2 (0.67)	1.6 (0.73)	2.2 (0.67)	1.4 (0.73)
t, p	1.04, 0.30	1.05, 0.30	0.03, 0.97	1.26, 0.21	0.52, 0.61	0.97, 0.34
Noun, sentence						
non-daily	2.1 (0.37)	2.1 (0.59)	2.1 (0.37)	2.1 (0.59)	2.1 (0.37)	1.4 (0.85)
daily	2.6 (0.53)	2.4 (0.53)	2.6 (0.53)	2.4 (0.53)	2.6 (0.53)	2.0 (0.71)
t, p	2.65, 0.03	1.60, 0.11	2.65, 0.03	1.60, 0.11	2.65, 0.03	2.12, 0.06
	phase-6		**phase-6**		**phase-6**	
Verb, sentence						
non-daily	1.6 (1.00)	1.9 (0.80)	1.8 (0.88)	1.9 (0.80)	1.7 (0.93)	1.8 (0.79)
daily	1.9 (0.93)	2.4 (0.53)	2.0 (0.87)	2.4 (0.53)	1.9 (0.93)	2.1 (0.78)
t, p	0.74, 0.46	1.81, 0.08	0.62, 0.54	1.81, 0.08	0.52, 0.60	1.02, 0.31
Adjective, sentence						
non-daily	2.2 (0.57)	1.3 (0.69)	2.5 (0.60)	1.3 (0.69)	1.7 (0.73)	1.3 (0.69)
daily	2.4 (0.53)	1.6 (0.88)	2.9 (0.33)	1.6 (0.88)	2.3 (0.71)	1.6 (0.88)
t, p	1.04, 0.30	0.83, 0.41	2.96, 0.01	0.83, 0.41	2.31, 0.02	1.18, 0.24
			phase-6		**phase-6**	
Preposition and connective, sentence						
non-daily	2.4 (0.70)	2.4 (0.65)	2.4 (0.70)	2.4 (0.65)	2.3 (0.77)	2.2 (0.71)
daily	2.8 (0.44)	2.8 (0.44)	2.8 (0.44)	2.8 (0.44)	2.6 (1.01)	2.6 (0.53)
t, p	2.42, 0.03	2.15, 0.05	2.42, 0.03	2.15, 0.05	0.99, 0.33	1.39, 0.17
	phase-6		**phase-6**			
Phrase						
non-daily	0.6 (0.68)	0.8 (0.95)	1.5 (0.60)	0.8 (0.95)	1.5 (0.60)	0.6 (0.85)
daily	0.4 (0.53)	1.1 (1.17)	1.4 (0.53)	1.1 (1.17)	1.3 (0.50)	0.9 (0.93)
t, p	0.53, 0.60	0.97, 0.33	0.34, 0.73	0.97, 0.33	0.62, 0.54	0.97, 0.33

Table 29: Paper 2 scores for parts of speech according to translation direction related to levels of phase-6 use

Data for School 2, showing high improvement also, was extracted for analysis. Totals for all data for this school are shown in the table following.

Total pupils	95
Primary users	10
Phase-6 users	44
Every day users	21
Once or twice a week users	11
Paper 1 correct answers	1532
Average Paper 1 correct answers	16
Paper 2 correct answers	3746
Average Paper 2 correct answers	39
Total difference between Papers 1 and 2 correct answers	2214
Average Paper 2 minus Paper 1 correct answers	23

Table 30: Test results for School 2

Pearson's correlations for data from this school were run for levels of phase-6 use against test paper results. These results are shown in the table following. For all tests, $N = 95$.

		Total results Paper 1	Total results Paper 2	Difference
Primary user	Pearson Correlation	.004	-.138	-.138
	Sig. (2-tailed)	.966	.182	.182
phase-6 user	Pearson Correlation	-.013	.084	.091
	Sig. (2-tailed)	.900	.417	.382
Every day user	Pearson Correlation	.041	.008	-.017
	Sig. (2-tailed)	.690	.938	.868
Once or twice a week	Pearson Correlation	-.083	.012	.061
	Sig. (2-tailed)	.425	.912	.554

Table 31: Levels of correlation between test results and levels of phase-6 use
(Note: * Correlation is significant at the 0.05 level (2-tailed). ** Correlation is significant at the 0.01 level (2-tailed).)

These tests do not reveal any highly significant correlations between results and background usage levels.

Mean average scores for results of the two papers, and for the difference in scores between papers, were calculated for four different groups for this school: all pupils who indicated they were every day users; all pupils who indicated they were not every day users; girls who indicated they were every day users; and girls who indicated they were not every day users. The mean average scores are shown in the table following. Highest mean average levels are highlighted in yellow.

GROUP	N	Average correct Paper 1	Average correct Paper 2	Average difference between papers
Every day users	21	16.57	39.57	23.00
Non every day users	74	16.00	39.39	23.39
Girl every day users	14	18.36	43.14	24.79
Girl non-every day users	37	16.49	40.62	24.14

Table 32: Average test results related to levels of use and gender in School 2

Girls who are every day users gained higher test results in Paper 1, maintained higher levels of results, but gained more than other groups by the time they took Paper 2. So girls who used phase-6 every day retained their higher performance, and produced results at the end of the test period that indicated they had gained more than had boys, or girls who did not use phase-6 every day. Would these girls have regarded these gains as arising from phase-6 use to any extent? It would certainly be interesting to know how important each of a number of contributory elements was in terms of the girls' perceptions of their successes in the second test: their willingness and perseverance; support from their teachers; support from their parents or those at home; and use of phase-6.

However, the two sets of analyses run at a specific school level suggest that phase-6 might be having an impact upon more able pupils. This could well be expected; pupils who are more able can use techniques that do not require the need for high levels of social interaction necessarily. So, phase-6 could be matching the learning approaches of this group of pupils, allowing them to explore vocabulary learning to take it to greater heights.

It is important to note that most of the analyses based on Test 1 and Test 2 data and improvements show that there was substantial difference between the schools. These were not simply a matter of the level of performance, as might be expected when schools have different cultures or intakes. The identified factors – translation direction, presence of context at test, part of speech - often interacted with school. Because phase-6 would be the same across schools, these interactions suggest that the schools might well be doing things differently from one another and that these differences might well be responsible for much of the story that the data might try to tell. Differences in practices encouraged by different teachers in the same school, as evidenced in Section 3.1, suggests that there are likely to be contributory factors that are pedagogically based.

5. THE RESEARCH STUDY IN THE SCHOOL IN CALIFORNIA

5.1 The focus for the study

This study was run by phase-6 and involved students being taught by two teachers within one school in California. The school is a 4-year high school, which prepares students for graduation to college (and also takes some students from a middle school on certain courses). For graduation, students are required to take a minimum of 2 years in a language, or a visual art, or a performing arts course. Most students who want to go to college stay at the school for 3 or 4 years, depending upon the college they want to go to. The nature of the school means that groups taught are completely mixed in terms of age, depending upon which year students choose to begin a course (and this includes a language study). Sometimes classes include students who might have failed a class on 1 or 2 previous occasions, and are attempting it for a second or a third time. Students who wish to take a language course at the school can take either French or Japanese in lieu of Spanish. There are 8 teachers in the school who teach Spanish.

Students in the study were aged between 13 and 17 years. Students had been using the phase-6 programme for only about 14 weeks when the test reported by teachers was run. It was unlikely, therefore, that any words would have moved into the sixth phase (and the numerical value shown in Column N in the spreadsheet gives some indication of this).

5.2 The study approach

It was not set up as a longitudinal study. The study results came from a single vocabulary test run in January 2008. The data were all collected on a single day. The aspect of learning tested was Spanish. Two teachers collected data:

- One teacher teaches Spanish 1 (at level 1). This class, in their first year of study in this subject (although some students might be repeating their first year of study), covered 5 chapters of the textbook, and 10 words were selected from each chapter, so students were tested on 50 vocabulary items. Half of this total (25 items) had been practiced with phase-6 before the test.
- The other teacher teaches Spanish 2 (at level 2, a class who have passed level 1). This class, in their second year of study in this subject (although some students might have had to repeat the first year of their study, or be repeating their second year of study), covered 4 chapters of the textbook, and 10 words were selected from each chapter, so students were tested on 40 vocabulary items. 20 of those items had been studied using phase-6, while 20 other items had not been integrated into phase-6.

5.3 The test items

All questions used in the tests were in the form of vocabulary items (words or phrases), given in English, which students had to write in Spanish. For ease and consistency, the teachers decided that to be "correct," the word had to be 100% correct. There was an even number of words chosen that were learned using phase-6 and not using phase-6. At the beginning of the year, the teachers chose 5 words from each chapter of the book to omit from entry into phase-6 for practice. When the teachers gave the assessment, they chose an equal number of words from each chapter that were in phase-6 and tried to choose words of equal difficulty. They chose all forms of words, tried not to use cognates, but tried to match verbs with verbs, adjectives with adjectives, and so on.

In the test with 50 questions, the balance of forms of words or phrases is shown in the table following.

Form of word or phrase	Number not practised in phase-6	Number practised in phase-6
Adjective	4	5
Adverb	2	1
Conjunction	1	1
Noun	6	7
Phrasal element	1	0
Phrase	4	3
Question	1	2
Verb	6	6

Table 33: Parts of speech involved in one of the tests related to phase-6 practice

5.4 Structure of the data within the spreadsheet

Within the 'Table' worksheet, the teacher's name is shown in column E. There are responses from 172 pupils in total. Column G shows the total number of correct responses for words learned using phase-6, while column I shows the total number of correct responses for words learned not using phase-6. Column K shows the numerical difference between columns G and I. Column N shows the level of score for words within phase-6, and uses a particular formula: 0 x p(1) + 0.5 x p(2) + 1 x p(3) + 1.5 x p(4) + 2 x p(5) + 2.5 x p(6), where p(y) is the number of words the pupil had in phase y at that point in time. Words in higher phases are weighted heavier, while words in phase-1 do not count at all. Column P shows the overall grade of the pupil in the class (not the grade of the test). Column Q shows any native speakers of Spanish (coded with a 1).

5.5 Questions for analysis of the responses

A number of key questions that might be posed of the data in the worksheet:

- How many native Spanish speakers are involved?
- What are overall indicators of outcomes when phase-6 is used, and when it is not used?
- Is there any indication that phase-6 is helping pupils in the tests?
- Is there any indication that levels of score within phase-6 are related to outcomes?
- Could any difference in outcome arise as a result of other influences (such as conscientiousness)?

6. FINDINGS FROM THE STUDY IN THE CALIFORNIAN SCHOOL

6.1 An overview of the study

This was not set up as a longitudinal study. The study was based on a single vocabulary test event run in January 2008. The data referring to "score" (activity level of a student) was collected on January 14, 2008. Two teachers collected data: a teacher of Spanish 2, during Periods 6 and 7, tested students based on 40 vocabulary items (20 of those items had been studied using phase-6, while 20 other items had never been integrated into phase-6; a teacher of Spanish 1, during Periods 0, 1 and 2, tested students based on 50 vocabulary items (25 of the items had been practiced with phase-6 before the test, while 25 had not used phase-6 for practice).

The analysis presented here explores individual student improvement, for instances where students were working with and without phase-6, where there was a threshold (such as more than 60% correct answers), students with high levels of phase-6 use compared to lower levels of phase-6 use (between group comparisons based on grade groups, activity level, and general comparison of usage levels). The study also explores combinations, within the same grade groups, and how activity level might have affected performance.

6.2 Correlation test results

A correlation test was run to see if score and activity measure might in any ways be related to correct answers when using phase-6 and not using phase-6.

		Number correct using phase-6	Number correct without phase-6
Score and activity measure	Pearson Correlation	.680(**)	.578(**)
	Sig. (2-tailed)	.000	.000
	N	172	172

Table 34: Levels of correlation between test scores and phase-6 use
(Note: ** Correlation is significant at the 0.01 level (2-tailed).)

Pearson's correlation test shows a high level of correlation at high levels of statistical significance between the measure of score and activity in using phase-6 and the numbers of correct responses, both using phase-6 and not using phase-6. However, the correlation between correct results and the score and activity measure is stronger when phase-6 is used (0.68 when used compared to 0.58 when not used). Subsequent analyses will explore in more depth whether this difference can be put down to phase-6 impact.

6.3 An overview of test results and improvements by group

In the original spreadsheet an error was involved in the calculating of percentage correct responses without phase-6 for students working at level 1. This error was corrected, and these analyses are based on corrected data.

Throughout these analyses the percentage correct, rather than number correct, has been used so that level 1 and level 2 students are measured on the same scale. Five classes participated, three studying first year Spanish and two studying second year Spanish. Ten students were native Spanish speakers; these were excluded from the analyses. Overall improvements by class are shown in the table following.

Level / Class	Class size	Excluding native Spanish speakers	Percentage improvement with phase 6			
			Mean	SD	Min	Max
1a	35	34	6.4	31.19	-60	83
1b	33	32	5.9	47.78	-67	200
1c	34	33	36.9	79.74	-50	400
2a	34	28	35.7	48.92	-36	200
2b	36	35	46.1	75.67	-67	400

Table 35: Improvement related to class and phase-6 use

Column O in the spreadsheet indicates those phase-6 users who are high-level users, and those who are lower level users. The analyses following are further limited to only those students who used phase-6 at a higher level (higher than the median value).

Level / Class	Class size	Excluding native Spanish speakers	Excluding students not using phase-6 at a high level	Percentage improvement with phase-6			
				Mean	SD	Min	Max
1a	35	34	18	11.7	20.96	-19	45
1b	33	32	10	6.1	26.63	-38	50
1c	34	33	14	41.7	36.87	-4	114
2a	34	28	13	41.1	39.72	0	143
2b	36	35	28	46.6	75.75	-13	400

Table 36: Improvement related to class and high levels of phase-6 use

Analyses are reported both on the full data set (excluding native Spanish speakers) and the subset identified as high-level users. In both sets of data, the average for each class showed improvement with the words involving phase-6 practice. The improvement was more substantial with the level 2 classes (average 41% and 47%), although one of the three level 1 classes experienced similar levels of improvement (average 42%).

6.4 Improvement using phase-6 and not using phase-6

Calculated improvements indicated that there was better performance on phase-6 words than on words not in phase-6. Although this may be due to use of phase-6, it might also be due to differences between the two sets of words. **Importantly therefore, further analyses here are based on the assumption that the word sets are equivalent. In the future, any studies conducted need to determine whether or not the word sets used are equivalent.** This might involve students at each level learning both sets of words without phase-6, or students at each level learning both sets of words with phase-6.

From the data in the Californian school, more students at level 2 benefited from phase-6 work than students at level 1, but there were consistent but small and non-significant benefits for phase-6 words even at level 1. Data were analysed separately for each class, enabling internal replication and some sense of reliability. Tests were one-tailed, asking whether using phase-6 improved performance.

Including all non-native Spanish speakers:

Level / Class	Eligible students (*N*)	Sign test		*t* test						
		Number better with phase-6	signif. level	with high levels of phase-6		With low levels of phase-6		*t*	signif. level	Cohen's *d*
				Mean %	SD	Mean %	SD			
1a	34	14	--	46	23.9	43	20.2	1.66	.053	0.14
1b	32	14	--	39	27.9	38	26.3	0.19	.425	0.04
1c	33	20	.148	38	24.0	31	21.5	**3.58**	**< .001**	0.31
2a	28	**25**	**.002**	64	20.7	51	19.7	**4.16**	**< .001**	0.64
2b	35	**30**	**< .001**	74	26.1	57	24.1	**7.02**	**< .001**	0.68

Degrees of freedom for the *t* tests are $N - 1$.
'—' indicates that the direction of the data were inconsistent with the one tailed test.

Table 37: Tests of significance on improvements related to class and phase-6 use for all students

Including only non-native Spanish speakers who were high level phase-6 users:

Level / Class	Eligible students (*N*)	Sign test		*t* test						
		Number better with phase-6	signif. level	with high levels of phase-6		With low levels of phase-6		*t*	signif. level	Cohen's *d*
				Mean %	SD	Mean %	SD			
1a	18	10	.408	60	20.4	55	18.2	**2.20**	**.021**	0.26
1b	10	5	.500	66	24.3	64	23.1	0.72	.246	0.08
1c	14	**12**	**.007**	57	19.6	44	21.1	**5.06**	**< .001**	0.64
2a	13	**12**	**.002**	74	20.0	55	20.0	**4.57**	**< .001**	0.95
2b	28	**25**	**< .001**	82.9	17.3	64	19.7	**6.84**	**< .001**	1.02

Degrees of freedom for the *t* tests are $N - 1$.
'—' indicates that the direction of the data were inconsistent with the one tailed test.

Table 38: Tests of significance on improvements related to class and phase-6 use for non-native speaking Spanish students

For level 2 classes, the benefits of phase-6 (or the use of easier words for the phase-6 set) appear to be substantial and robust. For level 1 classes there may be some benefit, but the picture is more mixed and less convincing.

This apparent interaction, between level of study and phase-6 usage (where level 2 students showed a greater difference between word sets than did level 1 students), is supported by ANOVA for the percentage improvement.

- For the data including all 162 students, the interaction is statistically significant, $F(1,160) = 31.96$, $MSE = .008$, $p < .001$, partial $\eta^2 = .16$ indicating a large effect (Cohen, 1988). The main effect of study level, where level 2 students recalled a greater percentage than level 1 students, was also substantial and significant, $F(1,160) = 37.55$, $MSE = .106$, $p < .001$, partial $\eta^2 = .19$ indicating a large effect. The main effect of word set, where the phase-6 words were better recalled than the other set of words, was also substantial and significant, $F(1,160) = 80.38$, $MSE = .008$, $p < .001$, partial $\eta^2 = .33$ indicating a very large effect.

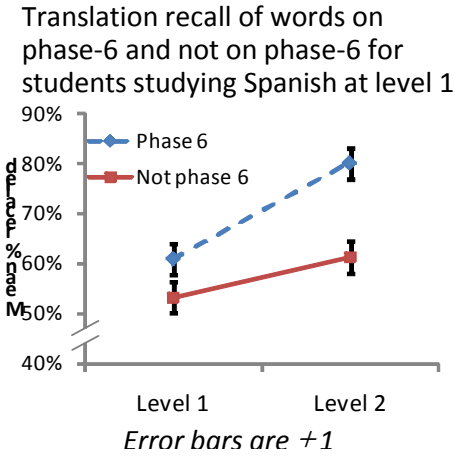


Figure 20: Differences between improvements related to level and phase-6

- For the data including students identified as high-level phase-6 users, the pattern is similar, with a larger effect of word set and a slightly smaller effect of study level. The interaction is statistically significant, $F(1,81) = 15.13$, $MSE = .009$, $p < .001$, partial $\eta^2 = .16$ indicating a large effect (Cohen, 1988). The main effect of study level, where level 2 students recalled a greater percentage than level 1 students, was also significant, $F(1,81) = 10.45$, $MSE = .073$, $p = .002$, partial $\eta^2 = .11$ indicating a medium sized effect. The main effect of word set, where the phase-6 words were better recalled than the other set of words, was also substantial and significant, $F(1,81) = 84.64$, $MSE = .009$, $p < .001$, partial $\eta^2 = .51$ indicating a very large effect.

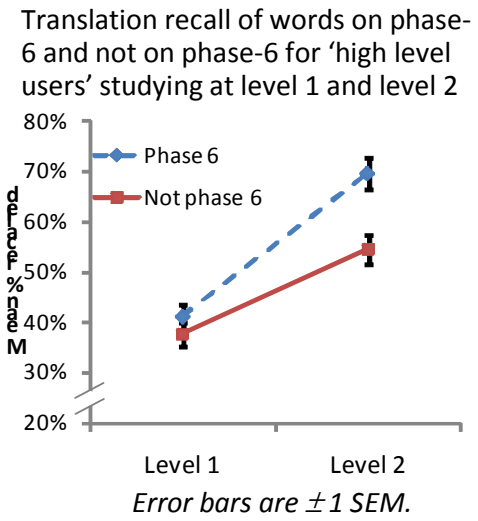


Figure 21: Differences between improvements related to low and high levels of use of phase-6

The results may be clear, but they are not as interpretable as one might like. The differences between the level 1 data and the level 2 data might be due to their level of learning (and one could construct a theory-based argument to support this) but they might also be due to the difference in the size of the to-be-learned set (50 for level 1, 40 for level 2) or the words in the set (the level 1 words might have been more challenging for level 1 students than were the level 2 words for level 2 students, or vice versa). The additional data suggested earlier – data on how well level 1 students learn the level 1 set and how well level 2 students learn the level 2 set without phase-6 involvement – would help to inform interpretation.

Performance while using phase-6 is logically a predictor of later performance on those words. Both the benefits of using phase-6 and the effort and ability that would lead students to make better use of phase-6 would lead to better performance on the final test.

The relationship between performance while using phase-6 and performance on the other set of words is potentially more complex. The use of phase-6 on some words may have led students to engage in somewhat similar practice on their own. So it is possible that phase-6 use might influence more effective study methods on the other words, thereby indirectly improving performance there. It is also likely that students' degree of effort and ability would influence both performance while using phase-6 and performance on all of the words, including the words not in phase-6.

Correlations between pairs of these three variables are reported below, both overall (for all students and for high-level phase-6 users) and by class.

All students

Level/ Class	*N*	phase-6 score		phase 6-test		Not phase 6 test		phase-6 and not phase-6 tests	phase-6 score and phase-6 test	phase-6 score and not phase-6 test
		Mean	*SD*	Mean	*SD*	Mean	*SD*			
1a	34	478	279.7	46	23.9	43	20.2	.90**	.66**	.66**
1b	32	332	255.5	39	27.9	38	26.3	.92**	.72**	.64**
1c	33	354	274.8	38	24.0	31	21.5	.89**	.73**	.62**
2a	28	405	208.5	64	20.7	51	19.7	.69**	.47*	.23
2b	35	543	240.9	74	26.1	57	24.1	.83**	.75**	.64**
Overall	162	425	263.8	52	28.5	44	24.1	.87**	.68**	.60**

** p < .001
* p < .05

Table 39: Levels of correlation by class between scores where phase-6 was used and where it was not used

High-level phase-6 users only

Level/ Class	*N*	score		phase-6 test		Not phase-6 test		phase-6 and not phase-6 tests	phase-6 score and phase-6 test	phase-6 score and not phase-6 test
		Mean	*SD*	Mean	*SD*	Mean	*SD*			
1a	18	668	249.2	60	20.4	55	18.2	.85**	.48*	.52*
1b	10	650	145.0	66	24.3	64	23.1	.87**	.14	.01
1c	14	613	207.9	57	19.6	44	21.1	.88**	.33	.44
2a	13	583	137.6	74	20.0	55	20.0	.72**	.17	.43
2b	28	636	153.5	83	17.3	64	19.7	.70**	.51*	.36*
Overall	83	632	182.5	70	21.8	57	20.9	.78**	.31*	.37**

** p < .001
* p < .05

Table 40: Levels of correlation by class between scores where phase-6 was used at a high level and where it was not used

Did using phase-6 improve students' performance on the test? Their performance could be predicted by:

- Their individual level of motivation and ability, which we can estimate by looking at their performance on the words that were not on phase-6.
- Their level (1st or 2nd year of study) and/or the difficulty of the set of words they studied.
- How well they scored while using phase-6.

A hierarchical regression was conducted to explore whether or not the phase-6 scores contribute significantly to predicting performance on the phase-6 words at test time.

- The first model removed the large amount of variance predicted by the test performance on the non-phase-6 words, $R = .87$, adjusted $R^2 = .76$, $F(1,160) = 498.54$, $MSE = .020$, $p < .001$

- The second model removed a significant further amount of variance as predicted by the students' level/word set, $R = .89$, adjusted $r^2 = .80$, $F(2,159) = 317.07$, $MSE = .016$, $p < .001$; for the change in the model $R^2 = .04$, $F(1,159) = 33.70$, $p < .001$.
- The third model showed that performance using phase-6 was a significant predictor of performance on the test for the phase-6 items, even after the variance shared with the first two steps were removed, $R = .92$, adjusted $R^2 = .84$, $F(3,158) = 276.00$, $MSE = .013$, $p < .001$; for the change in the model $R^2 = .04$, $F(1,158) = 39.66$, $p < .001$.

Thus, performance while using phase-6 makes a small but significant contribution to predicting performance on the test.

6.5 Future approaches

The analyses undertaken have highlighted a number of issues concerned with approaches to the study. A key question arising, where further evidence would enable more concrete conclusions to be drawn is: Were specific words counterbalanced in this research? The reason for needing to know the answer to this question is that, obviously, some sets of words will be easier to learn than other sets. Further studies need to draw on the methodologies that allow these effects to be identified and accounted for. When use of phase-6 is manipulated within student, as it was for these data, it is important that the two sets of words are varied across groups of students – so that the words studied with phase-6 include half of the words for any one student, but include all of the words across all students. A method that can be used to address this need is shown in the figure following.

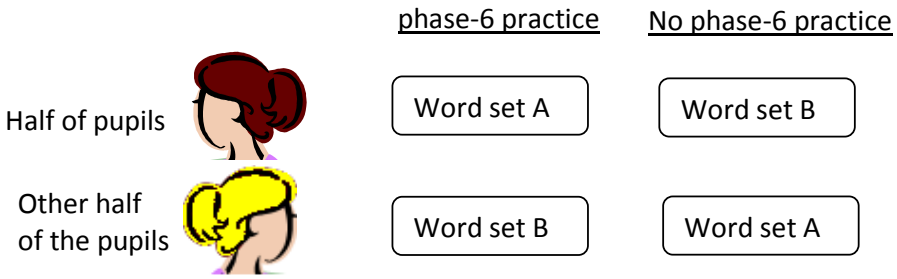


Figure 22: Method to address word equivalence level

Within this study, no evidence was available to indicate the sort of study or practice activity that was used for the other half of the words. In the future, this form of evidence will aid analysis and the drawing of conclusions from the data.

7. KEY FINDINGS AND CONCLUSIONS

7.1 Future approaches and methodologies

The two studies reported here have used different approaches. It is clear from the analysis of results from both studies that further studies in these locations can benefit from the experiences that have been gained from the findings. In particular, it is likely that higher levels of background contextual data would help. This should at least focus on gathering evidence about other forms of revision practice that are used by students when phase-6 is not used.

The study method in the school in California unfortunately did not account for the effect that a difference in difficulty of memorising sets of words might make. This report suggests a number of methods that would allow an analysis in future to account for this effect.

7.2 Differences across the two studies

There is an important difference between the two locations of the studies in terms of revision practice when not using phase-6. There is more use of boxes and cards divided into memory phases in schools in German schools. The study in schools in Germany was investigating to some extents the differences between memory practice using phase-6 (based on Ebbinghaus), and using box and cards (also based on Ebbinghaus), mixed with a variety of other short-term memorisation techniques. This is unlikely to have been the case in the school in California, where phase-6 (based on Ebbinghaus) was being compared to uses of word lists in books and the checking of each other's knowledge verbally (short-term memorisation practices not based on Ebbinghaus).

7.3 Main findings from the study in the German schools

It is perhaps, therefore, not surprising to find that phase-6 use levels are not showing overall strong benefits. However, if the tests are indicating potential differences between using technological versus non-technological revision processes, then they suggest that the technological process (phase-6) is certainly not worse than any non-technological process (and this is entirely consistent with all the evidence on other ICT forms of support).

There are indicators that pupils working at higher levels of performance may well be supported with the use of phase-6. It is entirely possible here that the technological process (phase-6) is providing a viable independent means of revision that allows these pupils to progress all the more rapidly. Other pupils working at lower levels of performance may well be supported by less independent approaches, involving higher levels of social interaction. This conjecture could be tested to some extent in future studies.

Arising from the study in German schools was a range of findings about the learning of language. Although these findings do not in themselves indicate anything general about the added value that phase-6 is bringing, they could be potentially valuable both to phase-6 (in terms of considering how content is structured in the future), and to schools (to alert them to pointers to support learning, and how phase-6 might interact effectively in certain places).

Key findings about language learning and revision approaches gained from the study included:

- The suggestion that once students have used phase-6, they see it as helpful and tend to use it again. Pupils' good intentions at the beginning of the year corresponded significantly with a greater likelihood of them actually using phase-6 during the year.
- School had a significant effect on pupils' initial performance and later performance.
- Girls translated significantly more of the items correctly from Paper 1 than did boys. The effect of gender was not significant at the end of the test period.
- As is usual when learning foreign vocabulary, pupils were significantly better at receptive translations than productive ones in both papers. This result may be an important conclusion for phase-6, in terms of the balance of access to receptive and productive vocabulary.
- Pupils translated items correctly significantly more often in Paper 1 and in Paper 2 when a sentence was present for context than when it was not. However, context provided a benefit for

verbs and adjectives, but not for nouns and prepositions/connectives in Paper 1, while in Paper context provided little benefit for nouns, but improved translations for the other parts of speech. This result may be important in terms of the forms of questions used within vocabulary training packages.

- Nouns were more often translated correctly in Paper 1 than the other parts of speech. This pattern is usual and is often attributed to the more concrete nature of many nouns. In Paper 2, prepositions and connectives were translated correctly almost as often as were nouns.
- On the first test, for strict scoring (and for flexible scoring to a smaller extent), there was a small but significant main effect of gender wherein girls produced more correct translations of phrases. On the second test, for strict scoring, the effect of gender observed in Paper 1 had been lost; girls produced roughly the same number of correct translations as boys with both scoring metrics.
- Based on large differences between schools, it would appear that other aspects of English language learning may well play a much larger role than has their use of phase-6.
- Two analyses run at a specific school level suggest that phase-6 might be having an impact upon more able pupils. This could well be expected; pupils who are more able can use techniques that do not require the need for high levels of social interaction necessarily. So, phase-6 could be matching the learning approaches of this group of pupils, allowing them to explore vocabulary learning to take it to greater heights.

7.4 Main findings from the study in the Californian school

The finding from the school in California, where an Ebbinghaus-style process is potentially being tested against non-Ebbinghaus process, offers an indication of impact, both at the level of improved performance, and of improved prediction. However, it is not possible to draw a firm conclusion that the differences in performance identified are due to phase-6 alone. Until further data is accessible about the comparative difficulty levels of the two sets of words, a final conclusion about impact cannot be stated.

7.5 Future studies

An ideal study to undertake to gather robust and rigorous evidence would be one which would generate pre- and post-test data on the learning of word or other subject matter, both using phase-6 and without phase-6 use at all, where teacher and learner backgrounds, methods and approaches would be otherwise similar.

A future study will need to gather both qualitative and quantitative evidence. These forms will allow levels of impact to be identified, but will also allow data to be gathered that will indicate reasons for impacts arising. Use of the phase-6 facilities is most developed currently in terms of language learning. It would seem appropriate in a future study to focus on the impact of phase-6 on language learning, in year 9, and its potential impact on choice of subject at GCSE subsequently. Within such a study, if students were generally taught with very similar materials (textbooks, laboratory tapes, etc.) and methods as those students using phase-6, then pre- and post-tests would be helpful in terms of data to evidence impact. It would then be possible to compare non-phase-6 words (from between the two groups on pre-tests), to show general equivalence. It would be possible to compare the post-tests on the non-phase-6 words to detect 'bleed over' effects from using phase-6 (if any occur) - but only if there is evidence that the teaching materials and methods are very similar to those used for the other students. If the teaching materials and methods are different, then those differences could cause or contribute to any differences observed between student sets.

From such a study, it would be possible to compare phase-6 words on the pre-test between user and non-user student sets, again to show equivalence. Comparing the two groups' post-test scores on these words would then be a good test of phase-6 effectiveness (again, if and only if, the teaching materials and methods were very similar).

The key of such a study is to provide evidence that students are similar in different sets and that the teaching materials and methods are similar, so that observed differences in post-test performance can not arguably be attributed to other differences between the groups, but to the difference of interest -

the use of phase-6. If it is possible to find equivalent classes that do not interact with one another - with matched teaching materials and methods - and ask one class to use phase-6 on half of a set of words and the other class not to use phase-6 at all, that would provide the most straightforward basis for drawing conclusions about what effects phase-6 use might be having. But in the same school, students may take a Spanish class at different times, but they would still interact with one another at other times. If they do interact, some students in the phase-6 group might introduce students in the control group to phase-6 and then the control becomes invalid.

So the best complete design probably involves pairs of schools that are well matched in terms of their student intake and their teaching materials and methods. This matching might be evidenced by showing similar test results in a previous year between a pair of schools and by describing the student demographics for each, and the teaching materials and methods used in each. Then students at one school would have phase-6 introduced for a set of words and would have another set of words that were not set up for phase-6 study. Students at both schools would take the pre- and post-tests on both sets of words. Then it would be possible to make comparisons that allow stronger inferences to be made about the effects of introducing phase-6. To remove the possibility of impact of a Hawthorn effect, it will be important that the study is set up so that data is gathered without influence of one school on the other, so that the school not using phase-6 is unaware of this practice if at all possible. In any pair of schools, teacher and learner backgrounds, methods and approaches should be similar other than the use of phase-6. It will then be possible to compare non-phase-6 words (from between the two groups on pre-tests), to show general equivalence.

For wider generalisation of results, it will important that paired schools cover different geographical areas, different locality settings (urban, rural and suburban settings), socio-economic settings, and banded, streamed or mixed groupings used for class teaching. Cohorts of some 100 students in each school would be ideal in terms of gathering robust data. Being able to gather data in written as well as recorded form would be clearly advantageous. This would especially useful if students were able to record and review their spoken responses against exact spoken responses.

7.6 Overview of findings

At this point the research indicates that:

- phase-6 might well be impacting upon the learning of words or phrases, but the evidence for this would need to be substantiated through a rigorous study designed according to control principles.
- The impact of phase-6 is likely to be affected greatly by pedagogical practices in schools.
- Findings suggest ways in which phase-6 could be used more effectively by learners and teachers, and suggests how the technological facilities could be developed to offer enhanced impact.

REFERENCES

Becta (2001a). *Primary Schools of the Future – Achieving Today*. Becta: Coventry

Becta (2001b). The *Secondary School of the Future – A Preliminary Report to the DfEE by Becta*. Becta: Coventry

Becta (2003a). *Primary Schools – ICT Standards. An Analysis of National Data from Ofsted and QCA*. Becta: Coventry

Becta (2003b). *Secondary Schools – ICT and Standards. An Analysis of National Data from Ofsted and QCA*. Becta: Coventry

Bjork, R. A. and Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn and R. Shiffrin (Eds.) *From learning processes to cognitive processes: Essays in honor of William K. Estes*. Hillsdale, NJ: Erlbaum.

Cox, M., Abbott, C., Webb, M., Blakeley, B., Beauchamp, T. and Rhodes, V. (2003a). *ICT and Attainment: A Review of the Research Literature. ICT in Schools Research and Evaluation Series No. 17*. Becta/DfES: Coventry/London

Cox, M., Webb, M., Abbott, C., Blakeley, B., Beauchamp, T., and Rhodes, V. (2003b). *ICT and Pedagogy: A Review of the Research Literature. ICT in Schools Research and Evaluation Series No. 18*. Becta/DfES: Coventry/London

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, 14, 215-235.

Cull, W. L., Shaughnessy, J. J. and Zechmeister, E. B. (1996). Expanding understanding of the expanding-pattern-of-retrieval mnemonic: Toward confidence in applicability. *Journal of Experimental Psychology: Applied*, 2, 365-378.

Fischer Family Trust (2003). *Impact of e-learning on GCSE results of 31,618 students, 2003*. Fischer Family Trust: Cardiff

Fischer Family Trust (2004). *Impact of e-learning on GCSE results of 105,617 students, 2004*. Fischer Family Trust: Cardiff

Fritz, C. O. and Morris, P. E. (2003). *Expanding retrieval practice: Investigating the parameters*. Paper presented to the Society for Applied Research in Memory and Cognition, 5th Biennial Conference, July 2-6, Aberdeen.

Fritz, C. O., Morris, P. E., Acton, M., Voelkel, A. R. and Etkind, R. (2007). Comparing and combining expanding retrieval practice and the keyword mnemonic for foreign vocabulary learning. *Applied Cognitive Psychology*, 21, 499-526.

Fritz, C. O., Morris, P. E., Nolan, D. and Singleton, J. (2007). Expanding retrieval practice: An effective aid to preschool children's learning. *Quarterly Journal of Experimental Psychology*, 60, 991-1004.

Harrison, C., Comber, C., Fisher, T., Haw, K., Lewin, C., Lunzer, E., McFarlane, A., Mavers, D., Scrimshaw, P., Somekh, B. and Watling, R. (2002). *ICT in Schools Research and Evaluation Series No. 7 – The Impact of Information and Communication Technologies on Pupil Learning and Attainment*. DfES: London

Landauer, T. K. and Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris and R. N. Sykes (Eds.) *Practical aspects of memory*. Academic Press: London

Leitner, S. (1972, 14th edition 1995). *So lernt: Der weg zum erfolg man lernen*. Herder: Freiburg

Morris, P. E. and Fritz, C. O. (2007). How to … improve your memory. *The Psychologist*, 19, 608-611.

Morris, P. E. and Gruneberg, M. M. (1996). Practical aspects of memory: the first 2500 years.

Morris, P. E., Fritz, C. O., Jackson, L., Nichol, E. and Roberts, E. (2005). Strategies for learning proper names: Expanding retrieval practice, meaning and imagery. *Applied Cognitive Psychology*, 19, 779-798.

NCET (1994). *ILS: Integrated Learning Systems. A report of the pilot evaluation of ILS in the UK*. NCET: Coventry

NCET (1996). *Integrated Learning Systems: A report of phase II of the pilot evaluation of ILS in the UK*. NCET: Coventry

Neuschatz, J. S., Preston, E. L., Toglia, M. P. and Neuschatz, J. S. (2003). A comparison of the efficacy of two name learning techniques: Expanding rehearsal vs. name-face imagery. Unpublished manuscript.

Ofsted (2001). *ICT in Schools: The Impact of Government Initiatives. An interim report April 2001*. Ofsted: London

Ofsted (2002). *ICT in Schools: Effect of Government Initiatives, HMI 423*. Ofsted: London

Ofsted (2004). *ICT in Schools: The Impact of Government Initiatives Five Years On, HMI 2050*. Ofsted: London

Passey, D. (2006). Technology enhancing learning: Analysing uses of information and communication technologies by primary and secondary school pupils with learning frameworks. *The Curriculum Journal*, 17, 2, 139 – 166.

Passey, D. and Rogers, C. with Machell, J. and McHugh, G. (2004). *The Motivational Effect of ICT on Pupils: A Department for Education and Skills Research Project 4RP/2002/050-3*. DfES: Nottingham

Pittard, V., Bannister, P. and Dunn, J. (2003). *The big pICTure: The Impact of ICT on Attainment, Motivation and Learning*. DfES: Nottingham

Reyna, V. F., Benbow, A. P., Boykin, A. W., Whitehurst, G. J. and Flawn, T. (2008). *Chapter 2: Report of the Subcommittee on standards of evidence*. Retrieved 3 October 2008 from the US Department of Education at www.ed.gov/about/bdscomm/list/mathpanel/reports.html

Roediger, H. L. and Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181-210.

Somekh, B., Barnes, B., Triggs, P., Sutherland, R., Passey, D., Holt, H., Harrison, C., Fisher, T., Joyes G. and Scott, R. (2001). *NGfL Research and Evaluation Series No. 2 – NGfL Pathfinders: Preliminary Report on the roll-out of the NGfL Programme in ten Pathfinder LEAs*. DfES and Becta: London

Somekh, B., Woodrow, D., Barnes, B., Triggs, P., Sutherland, R., Passey, D., Holt, H., Harrison, C., Fisher, T., Flett A. and Joyes G. (2002a). *ICT in Schools Research and Evaluation Series – No.10: NGfL Pathfinders Second Report on the roll-out of the NGfL Programme in ten Pathfinder LEAs*. DfES and Becta: London

Somekh, B., Woodrow, D., Barnes, B., Triggs, P., Sutherland, R., Passey, D., Holt, H., Harrison, C., Fisher, T., Flett A. and Joyes G. (2002b). *ICT in Schools Research and Evaluation Series – No.11. NGfL Pathfinders Final Report on the roll-out of the NGfL Programme in ten Pathfinder LEAs*. DfES: London and Becta: Coventry.

U.S. Department of Education, Institute of Education Sciences, and National Center for Education Evaluation and Regional Assistance (2003). *Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User Friendly Guide*. The Council for Excellence in Government: Washington DC

Underwood, J., Ault, A., Banyard, P., Bird, K., Dillon, G., Hayes, M., Selwood, I., Somekh, B. and Twining, P. (2005). *The impact of broadband in schools*. Becta: Coventry

Wood, D. (1998). *The UK ILS Evaluations: Final Report*. Becta: Coventry