REVIEW

December 2004

The effect of grammar teaching (sentence combining) in English on 5 to 16 year olds' accuracy and quality in written composition

Review written by the English Review Group

The EPPI-Centre is part of the Social Science Research Unit, Institute of Education, University of London

REVIEW GROUP

Authors and Review Team

Richard Andrews	University of York
Carole Torgerson	University of York
Sue Beverton	University of Durham
Allison Freeman	University of York
Terry Locke	Waikato University
Graham Low	University of York
Alison Robinson	University of York
Die Zhu	University of York

Contact details

Alison Robinson
Information Officer
Department of Educational Studies
University of York
York YO10 5DD

Email: ar31@york.ac.uk
Tel: 01904 433462
FAX: 01904 433459

ACKNOWLEDGEMENTS AND CONFLICTS OF INTEREST

The EPPI-Centre English Review Group and this review are part of the initiative on evidence-informed policy and practice at the EPPI-Centre, Social Science Research Unit, Institute of Education, University of London, funded by the Department for Education and Skills (DfES). Particular thanks go to Kelly Dickson, Diana Elbourne, Jo Garcia and all members of the EPPI-Centre team.

The Review Group acknowledges financial support from the DfES via the EPPI-Centre, via core institutional research funding from the Higher Education Funding Council for England; and from the Department of Educational Studies at the University of York. It is working within a University of York context where the National Health Service Centre for Reviews and Dissemination, the Department of Health Sciences, the Social Policy Research Unit and the Centre for Criminal Justice, Economics and Psychology are major players in evidence-informed research.

There are no conflicts of interest for any members of the group.

LIST OF ABBREVIATIONS

CT	Controlled trial
DES	Department of Education and Science (England and Wales)
DfEE	Department for Education and Employment (England and Wales)
PGCE	Postgraduate Certificate in Education
QA	Quality assurance
QCA	Qualifications and Curriculum Authority (England and Wales)
RCT	Randomised controlled trial

GLOSSARY

Accuracy of writing
Accuracy of sentence structure and correct use of punctuation with standard written English

Coherence
Relationships that link sentences together to form a meaningful flow of ideas or propositions. The links between sentences are often inferred, rather than explicitly flagged.

Cohesion
Grammatical or lexical (word-level) relationships that bind different parts of a text together: for example, 'however', 'on the one hand…', 'on the other hand…'.

Contextualised grammar teaching
Grammar teaching that takes account of the function of sentences and texts in context, and also of the relationship of sentences to higher (e.g. text) and lower (e.g. phrase, clause, word, morpheme [the smallest meaningful unit of grammar]) units of language description.

Decontextualised grammar teaching
Sometimes known as 'traditional' grammar teaching, this focuses on the internal dynamics and structure of the sentence or text, not on the context of written production (e.g. drill and practice).

Deep syntactic structures
These are the projected abstract, underlying structures of a sentence (as opposed to surface structures); more loosely, deep and surface structures form a binary contrasting pair of descriptors, the first being the supposed underlying meaning and the second the actual sentence we see or hear.

'Functional' grammar
The term used to describe Halliday's systemic-functional grammar (Halliday and Hasan, 1985). Such a grammar goes beyond the description or prescription or generation of sentences or texts. It aims to relate text and sentence to context and meaning.

Language awareness
An approach to teaching about language that aims to raise awareness of different aspects of language, as opposed to formal grammar teaching

Learning difficulties
General difficulties with learning, often assumed to face about twenty percent of the school population from time to time

Meta-language
A diction (specialised subset of language) used to discuss language, e.g. 'noun', 'syntax'

Oracy
The spoken equivalent of 'literacy'. The term is derived from an analogy with 'literacy'.

Paradigmatic
A set of linguistic items in which any member of the set can be substituted (grammatically) for another member. Paradigmatic items are in an 'or' relationship, whereas syntagmatic items (their opposite) are in an 'and' relationship to each other. For example, nouns and verbs each form a paradigmatic class.

Paragraph composition
Paragraphs have no grammatical status as such, but their arrangement within a text (e.g. 'the five-paragraph essay' in the USA tradition) is considered part of teaching textual grammar.

'Pedagogic' grammar
The distillation (usually of a traditional grammar) as used in textbooks for first or second language teaching

Punctuation
Surface markers for sentence structure, and/or, in the case of exclamation marks and question marks, indicators of tone and function

Quality of writing
Quality in terms of a set of criteria: for example, 'cohesion', 'imaginativeness', 'appropriateness of style', 'verve'. Usually judged inter-subjectively by a panel of experts (e.g. teachers).

Sentence combining
A teaching technique for linking sentences *horizontally*, i.e. not via their meaning or sub-grammatical character, but with connectives (e.g. conjunctions) or syntagmatically (see 'syntagmatic'). It can also cover sentence-embedding and other techniques for expanding and complicating the structure of sentences.

Sentence-diagramming
A technique deriving from structural and transformational grammars in which relationships between parts of a sentence are presented diagrammatically, often in tree-diagram form.

'Sentence' level grammar teaching
Teaching about the structural rules of sentence creation

Specific learning difficulties
Dyslexia and other specific difficulties with language learning

Syntagmatic
See 'paradigmatic'. Syntagmatic relationships can be conceived as in a chain or sequence: for example, the relationship between nouns and verbs in a sentence.

Syntax
Constraints which control acceptable word order within a sentence, or dominance relations (such as head noun + relative clause)

'Text' level grammar teaching
Teaching about the cohesion of a stretch of written composition. The term 'text grammar' applies the notion of grammar to whole texts, with an assumption of semantic (meaning), or pragmatic (meaning in use) coherence* (see 'coherence' above).

Text structure
Rules governing the internal arrangement of whole texts

Traditional grammar
Sentence grammars that tend to focus on the internal elements of the sentence, classifying 'parts of speech' and describing (and sometimes prescribing) the relationship between parts of speech.

Transformative/generative grammar
A transformative grammar attempts to systematise the changes that take place between the deep structures in language patterning and surface structures (i.e. the actual utterances made by speakers and writers); such a grammar is termed 'generative' because it is thought to be able to generate sentences or meaningful utterances, as opposed to merely describe or prescribe rules for their information.

Written composition
'Composition' is the term used to describe the putting together of words in an extended piece of writing.

This report should be cited as: Andrews R, Torgerson C, Beverton S, Freeman A, Locke T, Low G, Robinson A, Zhu D (2004) The effect of grammar teaching (sentence combining) in English on 5 to 16 year olds' accuracy and quality in written composition. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

TABLE OF CONTENTS

SUMMARY

Background

For over a century, there has been debate as to whether the teaching of grammar helps young people to learn to write well. The results have been inconclusive, partly because some parties in the debate have refused to acknowledge research evidence that suggests that the teaching of formal grammar (syntax, parts of speech) in a top-down approach is ineffective; partly because some of the research has been difficult to access and partly because previous studies and reviews have not been sufficiently comprehensive to answer the question of effectiveness conclusively. It is against this background that two in-depth reviews have been undertaken: one on the teaching of formal grammar (syntax), already published as Andrews *et al.* (2004) and the present review on the teaching of sentence combining.

Aim

The aim of the review is to shed conclusive light on the effect (or not) of teaching sentence combining on writing by 5 to 16 year olds in English.

Research questions

The overall research question for the systematic map of research is as follows:

What is the effect of grammar teaching in English on 5 to 16 year olds' accuracy and quality in written composition?

The specific research question for in-depth review in the present report is as follows:

What is the effect of teaching sentence combining in English on 5 to 16 year olds' accuracy and quality in written composition?

Methods

The systematic review (both the map and the in-depth study) used guidelines and tools devised by the EPPI-Centre (EPPI-Centre 2002a, 2002b and 2002c). In short, a protocol or plan for the research was drafted, including a provisional research question for the initial map of research in the field. Exclusion and inclusion criteria for the literature search were written. The protocol was peer-reviewed, revised and then published on the Research Evidence in Education website (http://eppi.ioe.ac.uk/reel). Research papers were searched, identified, screened for relevance and then keyworded to create an initial database. A map of research studies in the field was generated. From the map, two areas of research were identified for in-depth review: formal grammar (syntax) and sentence combining. Papers in this latter area were data-extracted and assessed for quality and weight of evidence with respect to the research question. A narrative synthesis of the results was produced.

Results

The initial electronic searching for research in the field since 1900 identified 4,691 papers, which were screened for potential relevance on the basis of title and abstract. A further 50 potentially relevant papers were identified through handsearching. A total of 267 papers were then obtained and re-screened against the inclusion/exclusion criteria on the basis of the full paper. Of these, 64 were found to meet the particular criteria for the review and constituted a map of the field. Twenty-six papers reported reviews and 38 reported primary research. Of the latter group, 20 papers, reporting on 18 studies, were deemed by the review group to be highly relevant to the in-depth review on sentence combining. Most of these studies (17) were from the USA; one was from Canada.

An overall synthesis of the results from the 18 studies examined in the in-depth review comes to a clear conclusion: that sentence combining is an effective means of improving the syntactic maturity of students in English between the ages of 5 and 16. All but two of the studies specify the age group they worked with: predominantly, this group ranged from fourth grade (9–10 year olds) to tenth grade (15–16 year olds), with the majority clustering in the upper years of primary/elementary schooling and the lower years of secondary schooling. The differences between the studies are largely inherent in the *degree* of advance that students learning sentence combining enjoy in terms of their syntactic maturity. In the most reliable studies, immediate post-test effects are seen to be positive with some tempering of the effect in delayed post-tests. In other words, as might be expected, gains made by being taught sentence combining in terms of written composition are greatest immediately after the intervention and tail off somewhat thereafter. Significantly, in the one study that undertakes a delayed post-test, syntactic maturity gains are maintained, albeit less dramatically than immediately after the event.

Conclusions

Taking into account the results and conclusions of the accompanying in-depth review on the teaching of formal grammar (Andrews *et al.*, 2004) the main implication for policy of the current review is that the National Curriculum in England and accompanying guidance needs to be revised to take into account the findings of research: that the teaching of formal grammar (and its derivatives) is ineffective; and the teaching of sentence combining is one (of probably a number of) method(s) that is effective.

In terms of practice, a very practical implication of the results of the present review is that it would be helpful if the future development of teaching materials and approaches included recognition of the effectiveness of sentence combining. There needs to be a review of the overall effectiveness of present materials designed to help young people to write; not all the practical suggestions put forward will be effective, and the emphasis on knowledge about language and language awareness, although useful and interesting in itself, may not be helping students to improve their writing skills.

In research terms, the present review(s) have achieved a ground-clearing operation, consolidating advances in the last 100 years or so, and mapping some of the territory for future research. It is suggested that further research needs to

move beyond studies of formal grammar and its effects on compositional skills; move beyond the USA into different contexts, taking into account the textual and contextual factors in learning to write; undertake some large-scale and longitudinal experimental studies to find out what works; improve the quality and reporting of such studies in the field and look at other ways of researching the effects, impact and nature of grammar(s) in learning to write.

1. BACKGROUND

1.1 Aims and rationale for current review

A systematic review is needed in order to ask the question: What is the effect of grammar teaching (sentence combining) on the accuracy and quality of 5 to 16 year olds' written composition?

The perennial question of whether grammar teaching helps writing quality and accuracy has haunted the teaching of English for over a century. Although there have been extensive reviews of the question (e.g. Macaulay, 1947; Wilkinson, 1971; Wyse, 2001), views remain polarised, with a belief among some teachers, newspapers and members of the public that such teaching is effective and among others that it is ineffective. A systematic review is therefore required to provide an authoritative account of the results of research into the question.

The English Review Group has already undertaken a systematic review of the effectiveness of grammar teaching (syntax) (Andrews *et al.*, 2004). Its review question is: What is the effect of the teaching of syntax on the accuracy and quality of 5 to 16 year olds' written composition? The question that drives the present review is related but different: What is the effect of the teaching of sentence combining on the accuracy and quality of 5–16 year olds' written composition?

The aim of the review therefore is to shed conclusive light on the effect (or not) of teaching sentence combining on writing by 5 to 16 year olds in English.

The objectives are as follows:

- to map the field of research on the effects of grammar teaching on writing in English-speaking countries for pupils aged between 5 and 16
- to undertake an in-depth review of one aspect of the field: the effect of teaching sentence combining on the quality and accuracy of 5 to 16 year olds' written composition

1.2 Definitional and conceptual issues

A very short history of grammar teaching: understanding the research context

We can divide the understanding of the nature of grammar, its place within language learning and the teaching of grammar within broad phases. Hudson (1992) suggests two phases to the understanding and teaching of formal written grammars.

According to Hudson, the first phase runs from 300 BC to 1957. This broad sweep of the history of grammars and grammar teaching has as its common strand the description of language and the subsequent prescription in 'grammar textbooks' in terms of how to write. The basic approach of these grammars is paradigmatic: that is, classes and categories of the language were defined and these were then taught as a means to write the language. In the Renaissance,

the principle of a scientific classificatory approach to written language gave rise to grammar in the curriculum (the other disciplines were Rhetoric and Logic, precursors to discourse analysis, mathematics and philosophy) and in turn to grammar schools. Grammar was often taught in this period via *progymnasmata* or exercises based on exemplary models of textual and sentence structure.

The publication of Chomsky's *Syntactic Structures* in 1957 marks the beginning of the second of these phases. Chomsky takes a structuralist approach, assuming that language can be described cross-sectionally or at any one moment in history in terms of a coherent system of rules. Such an approach is part of the tradition of cognitive neuroscientific theories of language production in that it is interested in the structural relationships between words, phrases and clauses in sentences, rather than in classificatory categories or 'parts of speech'. The quasi-mathematical formulae of Chomskian theory, with its distinction between *langue* and *parole* (between deep syntactic structures and surface manifestations in speech and in writing) gave rise to generative, transformational and later 'universal' grammars (see Damasio, 2000; Pinker, 1995). These grammars operated from basic principles in the construction of meaning and were intended to be able to generate intelligible sentences. Such generative capacity involved a *transformation* from deep structural rules and formulae to the actual utterances of everyday speech and writing.

Halliday approached language from a rather different perspective. One of his major contributions to the understanding of how language works was to combine the paradigmatic and syntagmatic. In his early work (summarised in Dixon 1965, pp 91–97), this complex relationship is couched in the primacy of form over context. In his later work – best interpreted in the early work of Kress (1994) – the relationship between form and context is explored in a more balanced way via the theory of systemic functional linguistics. A second major contribution by Halliday and his school, then, was to explore the relationships between the forms of language (e.g. lexical and syntactic elements) and the functions of language in particular contexts. The tradition of relating text to context (Fairclough, 1992; Halliday and Hasan, 1985; Hodge and Kress, 1993) sees grammatical knowledge as serving the development of critical understanding of how texts do their socialising work. Apart from the work on patterns of cohesion, including grammatical cohesion, in text (Halliday and Hasan 1976, 1985), one particular aspect of functional grammar which has been widely used in an educational context is 'grammatical metaphor' (Halliday 1988, 1989). This refers to the tendency of writers increasingly to reduce the number of words used as information becomes increasingly familiar. The final stages involve the condensation of entire events or processes into single abstract nouns. The point is that it is grammatical mechanisms which are used to downgrade, and hence to background, information. There are clear links here with Lakoff and Johnson's (1980) ontological metaphor, except that the functional approach covers a broader rhetorical context than simply treating events as entities or objects.

It is fully acknowledged in the present review that sentence level grammar is contingent upon the levels of text grammar ('above the level of the sentence') and word grammar ('below the level of the sentence'). Nevertheless, our aim will be to focus on sentence-level operations in teaching about writing and in learning to write.

Disillusion with traditional (that is to say, syntax-based and part of speech-based) grammar teaching required an alternative method of helping young people to

write. This emerged in the 1960s in the form of 'sentence combining', a generic term used to cover a range of practical methods for improving writing quality and accuracy.

Researchers, linguists, educators (but perhaps not policy-makers or the general public) had realized by the 1960s that the conventional approach to teaching syntax and the practice of 'parsing' (breaking down sentences into parts of speech in order to reconstruct them) was ineffective. Pointers toward sentence combining came from suggestions that studying formal grammar was less helpful than simply discussing grammatical constructions and usage *in the context of writing* (Harris, 1962; see also Calkins, 1980; Di Stefano and Killion, 1984) and that systematic practice in combining and expanding sentences could increase pupils' repertoire in syntactic structures, as well as improve the quality of their sentences (Hillocks and Smith, 1991).

The key source for much sentence-combining practice is O'Hare's *Sentence Combining: Improving Student Writing without Formal Grammar* (1973). The basic tenet of O'Hare's work is that written English is a dialect which is distinct from spoken English and that instruction should be based on language-learning techniques. The combining operation can be seen as a way of facilitating greater expression of ideas in various forms. Its success can be evaluated in terms of the length and complexity of sentences in pupil writing. The six factors that would be used to measure syntactic maturity are as follows:

- words per T-unit (a principal clause and any subordinate clause or non-clausal structure attached to, or embedded in it)
- clauses per T-unit
- words per clause
- noun phrases per 100 T-units
- adverbial clause
- adjectival clauses per 100 T-units

Basic to the practice of sentence combining, which can be defined as the manipulation of phrases and clauses to write more complex sentences, is that practice begins with a simple form, like a kernel or simple sentence. This kernel sentence is *combined* with another one (sentence combining), and/or elements are *embedded* into it. Combination can be effected via conjunctions or semi-colons or via subordination. A list of typical sentence-combining techniques would include the following:

- compounding sentences: for example, 'The bag felt heavy. It had lead in it.' becomes 'The bag felt heavy because it had lead in it.'
- compounding sentence elements
- subordinating one clause to another
- using appositives to connect ideas
- using participal phrases to connect ideas
- using absolute phrases to connect ideas

Embedding involves inserting into a simple sentence more complex constructions; for example, the simple sentence 'The bag felt heavy' could become, via a number of embeddings, 'The *blue* bag*, which had been lying on the platform overnight*, felt heavy *to Katya when she picked it up on that Thursday morning*.' One could embed yet further: 'The *faded* blue bag, which

(according to the detective) had been lying on platform *10* overnight, felt heavy to Katya when she picked it up *with difficulty* on that Thursday morning *in June*.' And so on.

Sentence combining (and embedding) became widespread in the 1970s and still has currency in the USA. At the 2004 American Educational Research Association Conference, for example, one of the papers (Saddler and Graham, 2004) 'examined the effects of sentence-combining practice on the writing and revising ability of forty-three fourth grade students', and found that it increased oral and written sentence-combining skills for the experimental group, as well as improved the quality of the revised stories for the group.

Key definitions

Grammar refers, as far as the present project is concerned, to written sentence and text grammar. It includes the study of syntax (word order), clause and phrase structure, and the classification of parts of speech (e.g. noun, verb, predicate, clause, etc.); and issues regarding the cohesion and coherence of whole texts. It can be both descriptive, in that it describes the existing patterns of sentences and texts; and, in sentence terms, also generative or transformative, in that rules can be defined which can generate grammatically acceptable sentences (the transformation being from basic rules through to actual sentences). Studies of words or sub-components of words are not part of the study of grammar *per se*. Similarly, studies in language awareness are not, strictly speaking, part of the present review, although the larger category of language awareness may come into play in considerations of grammar.

By *written composition*, we mean extended pieces of writing (in handwriting, in type or via word-processing) in a variety of genres or text-types.

In focusing on *accuracy*, we mean to place emphasis on appropriateness of grammatical form for particular purposes. We are not concerned with spelling accuracy, neither with legibility, neatness of handwriting or vocabulary (except where it bears upon sentence grammar). The emphasis on *quality* is there to distinguish our study from an interest in *quantity*.

By *English-speaking countries*, we mean countries where English is spoken as a first language by a significant segment of the population[1]. We include UK, Ireland, USA, Canada, Australia, New Zealand, Jamaica and other countries in the Caribbean, Gibraltar and South Africa; we exclude India, Pakistan, Hong Kong, Singapore, Bangladesh, Malaysia and China.

[1] We are aware that some detailed awareness needs to be demonstrated on schools' ethnic composition when studies are data-extracted. EPPI Reviewer enables us to record the ethnic composition of classes and this is a factor we shall take into account in our narrative synthesis of results. We shall inevitably be constrained by whether research studies report on the ethnic composition of the classes they investigate.

1.3 Policy and practice background

The teaching of grammar: the policy, practice and research contexts

Since the publication of the Kingman Report (DES, 1988), there has been a conviction amongst curriculum writers and policy-makers in England that grammar teaching to young learners of English is a good thing; that it will improve their written English and their ability to talk about language; that talking about language is helpful in understanding language and, in turn, in improving its use; and that such reflection and discussion *about* language should start earlier than had previously been thought possible or desirable.

It should be said at the start that such a conviction flies in the face of research evidence. Perera (1984, p 12) notes:

> Since the beginning of the [20th] century, a body of research has accumulated that indicates that grammatical construction, unrelated to pupils' other language work, does not lead to an improvement in the quality of their own writing or in their level of comprehension. Furthermore, the majority of children under about fourteen seem to become confused by grammatical labels and descriptions. It is obviously harmful for children to be made to feel that they 'can't do English' because they cannot label, say, an auxiliary verb, when they are perfectly capable of using a wide range of auxiliary verbs accurately and appropriately. There is a brief summary of this research evidence in Wilkinson (1971, pp 32–35).

Wilkinson notes that, although grammar is a useful descriptive and analytical tool, 'other claims made for it are nearly all without foundation' (ibid, p 32). Studies in the twentieth century have suggested that the learning of formal, traditional (i.e. not transformative) grammar has no beneficial effect on children's written work (Rice, 1903); that training in formal grammar does not improve pupils' composition (Asker, 1923; Macaulay, 1947; Robinson, 1960); that ability in grammar is more related to ability in some other subjects than in English composition (Boraas, 1917; Segal and Barr, 1926); that a knowledge of grammar is of no general help in correcting faulty usage (Benfer, 1935; Catherwood, 1932); that grammar is often taught to children who have not the maturity or intelligence to understand it (Macaulay, 1947; Symonds, 1931); and that teaching grammar may actually hinder the development of children's English (Macaulay, 1947).

A recent 'critical review' of the 'empirical evidence' on the teaching of grammar provides an overview of research studies in English-speaking countries (Wyse, 2001). This review concludes that 'the teaching of grammar (using a range of models) has negligible positive effects on improving secondary pupils' writing' (p 422).

Policy and practice in the 1970s and 1980s in England has followed a line characterised by the Bullock Report (DES, 1975); specifically that it was *teachers* who needed to know about grammatical construction so that they could understand pupils' writing problems and intervene accordingly and appropriately:

> We are not suggesting that the answer to improved standards is to be found in…more grammar exercises, more formal speech training, more comprehension extracts. We believe that language competence grows incrementally, through an interaction of writing, talk, reading and

> experience, the body of resulting work forming an organic whole. But this does not mean it can be taken for granted, that the teacher does not exercise a conscious influence on the nature and quality of its growth (pp 7–8).

In New Zealand, recent emphasis (Ministry of Education, 1996) has been on knowledge about language and exploring language rather than on grammar teaching *per se*. There is scepticism about the value of grammar teaching for the improvement of writing ability (Elley *et al.*, 1979, p 98):

> The primary purpose of this investigation was to determine the direct effects of a study of transformational-generative grammar on the language growth of secondary school pupils. The results presented show that the effects of the three years of such grammar study are negligible. Those pupils who studied no formal grammar for three years demonstrated competence in writing and related language skills equal to that shown by the pupils who studied transformational or traditional grammar. Furthermore, their attitude to English as a subject of study was more positive.

In these respects, the English and New Zealand positions are similar: they have seen a diffusion of emphasis on grammar teaching and a resultant reorientation around language awareness. *Exploring Language: A Handbook for Teachers* (Ministry of Education, 1996, p 3) states:

> Knowledge of the workings of language is also essential for teachers to be able to examine and assess their students' language use in a systematic and productive way. Behind messy handwriting and creative spelling, there could well be signs of interesting language development and attempts at new complexities and variation that could pass unnoticed by those who do not have a knowledge of understanding to recognize them. How can a teacher appreciate a student's new developments with passive verbs or modal auxiliaries if these concepts themselves are not known or recognized?

More recently, in England and Wales, the National Literacy Strategy (which operated for 7 to 11 year olds from 1997 before being extended to 11 to 14 year olds in 2002) has issued a book and video entitled *Grammar for Writing* (DfEE 2000), aimed particularly at the teaching of 7 to 11 year olds. The basic principle behind this relatively recent initiative is that 'all pupils have extensive grammatical knowledge' (p 7) and that teaching that focuses on grammar helps to make this knowledge explicit; this, the book and video argue, helps to improve young people's writing through providing them with an increase in 'the range of choices open to them when they write' (ibid). Throughout, there is a distinction between spoken grammars and written grammars, and a clear objective to support the development of a command in sentence construction. In pedagogic terms, the emphasis of the book is on teaching at the point of composition, rather than correcting after the event. While eschewing a return to the descriptive and prescriptive grammar teaching of the 1950s and 1960s, this approach does focus clearly on the improvement of sentence structure and uses extensive 'knowledge about language' and increased language awareness as a means to help pupils to write better English. It consists of a detailed programme for using sentence grammar to improve sentence construction, via explicit teaching. As such, it represents a middle ground between traditional grammar teaching on the one

hand, and language awareness arising from the use of language in speech and writing on the other.

Interestingly, the National Literacy Strategy in the UK, in its publication *Grammar for Writing* (see DfEE, 2000) does not explicitly mention sentence combining. The two key recent papers from the Qualifications and Curriculum Authority (QCA) – *The Grammar Papers* (1998) and *Not Whether But How* (1999) – inform the debate about grammar teaching and provide some very useful instances of how grammar has been used, or could be used, to enliven English teaching. They appear to accept the argument that the teaching of formal grammar to improve writing quality and accuracy has been lost, but look for ways in the English curriculum in which learning about language can inform pupils' language development.

Whose particular grammar are we concerned with?

The National Curriculum for England and Wales, when it was first established in the late 1980s and early 1990s, indicated that children should be able to talk about 'grammatical differences between Standard English and a non-standard variety'. Specifically, 'Standard English' refers to a broad set of conventions observed in the UK about the use of written English. Such a conception is not affected by accent. You can speak standard spoken English with a Scottish accent and written standard English is even less culturally specific. However, it has to be acknowledged that written American English has a different grammar from written British English and that pronunciation can impact on unstressed morphemes: that is, parts of words that carry relatively weak emphasis differently according to pronunciation and that sometimes affect meaning as a result.

Even with a broadly accepted set of conventions, there is room for disagreement and variation. 'The heading' for this section finishes with a preposition: 'are we concerned *with*'. Some people would find such a construction unacceptable and would rather see it expressed in writing as 'with whose particular grammar are we concerned?' Such variations tend to come down to questions of taste and preference. We could argue that the former is clearer, more elegant and more colloquial (and therefore more readily comprehensible); but someone else might argue that our version is less elegant, less formal (it is certainly less formal) and less *appropriate* than the latter version. Opinions about the nature of grammar, grammatical 'correctness' and the teaching (or not) of grammar make this a contentious field.

Hudson's (1992) book, *Teaching Grammar*, suggests that 'until you know what is on the menu you can't choose from it' (p. xi). In arguing the case for increased awareness of language construction amongst teachers, he is saying something similar to the Bullock Report's position that it is useful for teachers to know about grammatical construction so that they can help pupils appropriately, or Perera's (1984) similar conclusion. It may be that there is a degree of consensus among researchers and policy-makers from the 1970s to the 1990s; specifically, that, at the very least, teachers of English should know about grammar so that they can advise their pupils according to their particular needs. Perhaps a key distinction to be made at this point – one that might have a bearing on the systematic review undertaken – is how much teachers need to know about grammar in order to teach writing, and how much pupils need to know in order to write well.

Kress (1994) provides another, more radical perspective on grammar and grammar teaching. He starts from the premise that a grammar 'is adequate if that grammar allows a speaker to express the range of meanings which that speaker needs to express in such a way as to be understood in a regular and predictable manner by a fellow user of that grammar' (p 160). In other words, a grammar is an agreed set of conventions for a particular social group or in a particular social situation; it is not a Chomskian 'universal grammar'. Thus a child's grammar may differ from an adult's and 'the whole idea of correcting a child's grammar assumes that the child's grammar is inadequate to the expression of the child's meanings' (p 163).

Which grammar?

The developers of *Exploring Language* (Ministry of Education, 1996) assert that 'students and teachers need to be able to use a nationally agreed metalanguage of concepts and terminology to describe and discuss language' (p 7). In describing the process they went through to decide on this nationally agreed metalanguage, they write, 'rather than subscribing to one particular school of thought or approach to describing language, this book uses the descriptions and terminology that *will be most useful to teachers in the work with students*' (our italics). They describe this approach as eclectic. It could be argued that the writers of the book favoured Quirk *et al.* (1985) – a descriptive approach to grammar – over systemic functional grammar as the basis for their taxonomy, and therefore that they opted for a bottom-up grammar: one that does not deal with such aspects as cohesion or coherence. There is clearly a metalanguage set out in *Grammar for Writing* (DfEE, 2000), mentioned in the previous section.

Our own position in the current review is to be open to both the bottom-up approach and to the top-down approach in the systematic map of the research in the field and then to focus on sentence grammar (sentence combining) for the in-depth review. In the former case, the constructions and choices made are *informed* by semantic, textual and contextual factors. In the case of contextual factors, there is an emphasis on parts of speech and combining rules without much consideration of why certain combinations are acceptable and others not.

Grammar and the National Curriculum

The Kingman Report (DES, 1988), mentioned earlier, was a key document in the formulation of policy on grammar teaching and language awareness in England and Wales. Its general recommendations were to increase language awareness among pupils by increasing it among teachers at both primary and secondary levels in schooling. Although one of its recommendations – that 'by the end of the [20th] century a prerequisite for entry to the teaching profession as an English specialist should normally be a first degree which incorporates the study of both contemporary and historical linguistic form and use' (p 70) – has not been met, the advent of English Language courses at Advanced Level and the development of the National Literacy Strategy are indications of an increased emphasis on language study.

The study of grammar – the forms of the language at sentence and discourse levels – is but a part of the model proposed by Kingman, which also includes three other dimensions: communication and comprehension, acquisition and development, and historical and geographical variation (ibid, pp 17ff).

The latest version of the National Curriculum for England suggests that 'pupils should be taught some of the grammatical features of written standard English' as early as Key Stage 1 (ages 5 to 7) (DfEE 1999, p 21). By Key Stage 2 (ages 7 to 11), as far as reading is concerned and under the heading of 'Language structure and variation':

> To read texts with greater accuracy and understanding, pupils should be taught to identify and comment on features of English at word, sentence and text level, using appropriate terminology (p 26).

One example is the use of varying sentence length and structure. In writing, at this stage, 'some of the differences between standard and non-standard English usage, including subject-verb agreements and use of prepositions' (p. 29) should be taught. More detail is forthcoming on language structure, where pupils should be taught:

- word classes and the grammatical functions of words, including nouns, adjectives, adverbs, pronouns, prepositions, conjunctions, articles
- the features of different types of sentence, including statements, questions and commands, and how to use them
- the grammar of complex sentences, including clauses, phrases and connectives (p 29)

The refinement of these details at Key Stages 3 and 4 (11 to 16) simply requires that pupils should be taught 'the principles of sentence grammar…and use this knowledge in their writing'. Such teaching should include 'word classes or parts of speech and their grammatical functions' and 'the structure of phrases and clauses and how they can be combined' (p 38).

It is interesting to note that the major push on grammar teaching comes at Key Stage 2 (7 to 11). Wyse (2001) argues that the 'Grammar for Writing' initiative is insufficiently supported by empirical evidence on the teaching of grammar 'and that changes will need to be made to English curriculum policy and pedagogy if children's writing is to further improve' (p 411). The debate continues.

1.4 Research background

The first major study of the use of formal grammar in the teaching of writing was that by Macauley (1947). However, Macauley's study focused on the question of at what stage formal grammar should be taught, rather than whether it was appropriate and effective for it to be taught. He came to the conclusion, after a number of tests on the effectiveness of grammar teaching, that neither upper primary (i.e. 11–12 year old) pupils nor junior secondary (i.e. 13–14 year old) pupils could be depended on to recognise simple examples of nouns, verbs, pronouns, adjectives or adverbs after several years of having been taught it in English lessons (the latter group, for six years). Only upper secondary (i.e. 15–17 year old) pupils and those in the top boys' and girls' classes in each year were able to reach the 50% pass standard set in Macauley's tests. His overall conclusions are that scores rise with age and schooling but that, for most pupils, age and schooling are not in themselves enough for a mastery of even the most simple rules in English formal grammar and that 'those who pass our standard are few in number and are in the best of the [upper] secondary classes' (p 162). The implications Macauley draws out for the stages of schooling are clear: there

is no point in trying to teach formal grammar in the primary years or even in the lower secondary years; it is a practice and field best reserved (if at all) for brighter pupils in the last years of secondary schooling. The study does not look at the effect of such teaching on writing accuracy or quality, but it does point out the difficulties of the first part of our research question: the teaching (and by implication, the learning) of formal grammar.

As Braddock *et al.* (1963) note, in a review of the state of knowledge about composition for the National Council of Teachers of English (USA), the merits of formal grammar as an instructional aid is 'one of the most heavily investigated problems in the teaching of writing' (p 37). They summarise the field by stating that 'study after study based on objective testing rather than actual writing confirms that instruction in formal grammar has little or no effect on the quality of student composition' (p 37) and that 'direct methods' rather than methods based upon a knowledge of so-called related grammatical elements are more likely to be effective.

A particularly significant study undertaken in the UK was that by Harris (1962), which compared the effect of instruction in formal grammar and functional grammar over a period of two years on the writing of 228 London pupils aged 12 to 14. This study has been seen as significant because of its longitudinal dimension and its comparison of formal grammar teaching on the one hand, and 'functional or 'direct' (i.e. no formal grammar teaching) on the other.

Harris writes in the abstract to the thesis:

> In this work, the value of the traditional English grammar lesson in helping children to write correctly was tested. The grammar lesson was found to be certainly not superior, and in most instances was inferior, to direct practice in writing skills. The progress of five forms having no grammar lesson was measured on eleven counts against that of five similar forms following the same English course but taking one lesson a week of English grammar. At the end of two academic years, of the fifty-five resultant scores, twenty-five proved highly reliable.

Eleven measures were used in judging essays written at the beginning and end of the experimental period. These were the average length of correct simple sentences (not reliable); instances of omission of the full-stop (fairly reliable); the number of words per common error (very reliable); the variety of correct sentence patterns used (very reliable); the number of correct non-simple sentences minus correct simple sentences (fairly reliable); the total number of subordinate clauses (very reliable); the total number of words (not reliable); the number of correct complex sentences minus the number of incorrect (very reliable); the number of correct simple sentences with two modifying phrases (fairly reliable); the number of total correct sentences minus incorrect (fairly reliable); and the number of adjectival phrases and clauses (fairly reliable). There were thus five very reliable measures, four fairly reliable ones and two not reliable.

Detailed results show that in ten out of the 25 very reliable scores, significant gains were made by the non-grammar classes (n = 109), with no significant gains being made by the classes studying grammar (n = 119). Specifically, 'mechanical, conventional correctness – as in the number of words per common error; maturity of style – as in the variety of sentence patterns used; the control of complex

relationships – as in the number of correct complex sentences; as well as general overall correctness, seen in the total number of correct sentences, were all improved significantly in groups practising direct writing skills as compared with the groups studying formal grammar' (p 203).

Harris is aware that the results must be treated with caution because the experimental and control groups were not strictly comparable. However, he claims that there was no critical need to equate exactly the groups in each school; that both the general and English attainment 'were roughly of the same standard' (p 206); and that the content and order of the grammar and non-grammar syllabi were not significant 'since formal grammar itself has a vague and fluctuating meaning in present usage' (p206).

At the time the thesis was written – and we can safely assume, for the decade or so prior to its writing – about one-fifth of English class time was devoted to the teaching of 'formal' grammar. This figure is reflected in the amount of space given to grammar instruction and exercises in textbooks at the time. Harris questions, in the light of his findings, whether such time is worthwhile, particularly as his results echo those of Macauley in that 'no real likelihood exists of successfully teaching formal English grammar to any but bright children' (Harris, 1962, p 196).

Harris therefore argues for a 'grammar of situation': that is, the study and practice of language in action rather than of the artificially narrow formal grammars.

What are the limitations of Harris' study? First, although the empirical data-gathering part of the study took place over two years, Harris admits himself that this is the 'source…of much of the organisational fallibility' (p 111). Second, there were only two forms running in pairs in each school, and thus the sample is relatively small. Third, it was not possible to have complete control over the experimental situation over a two-year period: 'A number of variables had to be accepted without adequate control, in the hope that the difference between the work done by the experimental groups would be sufficiently large and clear to counter-balance in the results uncertainty due to uncontrolled variables or to lack of random or representative sampling' (p 112). Because the five schools used in the study consisted of two grammar schools, two technical/comprehensive and one secondary modern, the schools 'necessarily decided the groups of children who could be used, and in this there was no possibility of selecting two ideally equated groups, either in intelligence, background or attainment' (p 113). In other words, although every effort was made to control the study (for example, in one teacher teaching both the control and experimental groups in each of the schools), there were variables that were not controlled. The results of the study, therefore, have to be taken with a degree of caution.

Braddock *et al.* (1963) point out that the Harris study 'does not necessarily prove…the ineffectiveness of instruction based on structural or generative grammar' (p 83).

Tomlinson (1994) is the most critical of Harris' approach. He points out the fact that the study sample was neither randomised nor fully controlled, but accepts that such weaknesses were not decisive. More important for Tomlinson is the fact that there seems to be no clear distinction in the Harris study between the two types of grammar being taught: on the one hand, formal teaching of grammar (or indeed, teaching of formal grammar); and, on the other, what appears to be more time devoted to composition but with coaching in error avoidance – what might be

described as ‘a linguistically informed process of teaching composition’. The fact that the same teacher taught both experimental and ‘control’ classes in a single school suggests to Tomlinson that the ‘non-grammar’ class probably was in receipt of indirect grammar teaching rather than no grammar teaching. Tomlinson argues that the over-simplification of Harris results and conclusions led to an uncritical acceptance that grammar teaching (i.e. formal, ‘arid’, ‘parts of speech’ grammar) was unproductive, and thus to policy and practice decisions that were based on a simplistic distillation of research that was itself flawed in two important respects.

Wyse (2001) defends Harris against Tomlinson’s criticisms that his distinction between ‘grammar’ and ‘non-grammar’ approaches was really a distinction between a formal grammar approach and an informal grammar approach; we agree with Wyse that such a point does not invalidate Harris’ findings. However, we do have to accept that the Harris study was not entirely reliable.
What is interesting is how policy and practice tend to over-simplify the results of research according to the *zeitgeist* or the biases of the period. Such a phenomenon suggests that there needs to be better summarised reporting of research, with implications for policy and practice drawn out to help define exactly what these implications might be.

Two previous systematic reviews have been published in the field (Hillocks, 1984, 1986; Asher, 1990).

In 1986, Hillocks published a meta-analysis of experimental studies designed to improve the teaching of written composition. He analysed the experimental research between 1960 and 1982 on all interventions to improve written composition through a series of meta-analyses. Two of these were meta-analyses of trials of the effect of teaching grammar and sentence combining. Hillocks concludes that grammar instruction led to a statistically significant decline in pupil writing ability and that this was the only instructional method of those examined not to produce gains in writing ability. Five experimental/control treatments focused on grammar in one treatment but not in the other. When compared with courses designed to teach writing tasks directly, the grammar group performed consistently worse on the essay writing exercise. The mean effect size (a given treatment gain or loss expressed in standard score units) for grammar instruction was –0.29 (CI –0.40 to –0.17). Hillocks concluded that ‘every other focus of instruction examined in this review is stronger’ (1984, p 160). Five studies were included in the meta-analysis that focused on sentence combining as a method of instruction. The mean effect size for sentence combining was 0.35 (CI 0.19 to 0.51, statistically significant positive effect). Hillocks concludes that his research shows ‘sentence combining, on the average, to be more than twice as effective as free writing as a means of enhancing the quality of student writing’ (1984, p 161). However, Hillocks was comparing the pooled effect sizes calculated in the meta-analyses for various interventions versus control groups, rather than pooled effect sizes for grammar interventions compared directly with other interventions.

In 1990, Asher updated Hillocks’ 1984 systematic review on the effectiveness of sentence combining on writing. He found six experimental studies reported after 1983 and added them to the Hillocks’ original 21 studies (Hillocks used only five studies in his sentence-combining meta-analysis). With all these 27 studies, Asher calculated effect sizes for different age ranges:

Elementary school students: 0.34
Junior high students: 0.90
Senior high students: 0.08
College freshmen: 0.26

Asher concludes that '…sentence combining works best at the Junior High School level, but its effectiveness drops both before and after that level…' (p 152).

This present systematic review is, therefore, required because the only other systematic reviews in the field are now fifteen and twenty years out of date, and because these reviews didn't focus exclusively on investigating the effectiveness of grammar teaching on the quality of children's and young people's (aged between 5 and 16) writing, but rather included other populations, in particular 'college students'.

1.5 Authors, funders and other users of the review

The authors of the present review are stated at the beginning of the report. They include researchers and a doctoral student from the Department of Educational Studies at the University of York. Two of the researchers are ex-Heads of English in secondary schools in the UK. One is an applied linguist. Additionally, there are researchers from Durham (UK) and Waikato (New Zealand) Universities, one of whom has held senior posts in primary education and the other in secondary education. Furthermore, there is an experienced information officer on the review team.

The review has been funded by the Department for Education and Skills via the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) at the Institute of Education, University of London) and by the Department of Educational Studies at the University of York.

The Department of Educational Studies has been developing its links with schools interested in research since 2003 (see Department Plan, available from Alison Robinson). Such links will enable more teachers than those on the advisory group to comment on, contribute to and disseminate the work of the English Group. In addition, following a meeting with the Teacher Training Agency and PGCE students in June 2003, PGCE tutors and students will be involved in a pilot project to write summaries of the present research review (and previous reviews) and to prepare sample lessons arising from the research findings. In addition, a pupil from a secondary school has written a pupil summary of the final review. The dissemination strategy of the English Review Group was discussed at the steering group meetings in September 2003, February 2004 and September 2004.

Users summaries of the review(s) will be written by teachers, teacher educators, students, governors and policy-makers. Representatives from each of these constituencies (except pupils) have contributed to the direction and design of the review through the English Review Group's advisory steering committee.

1.6 Review questions

Research question for systematic map

What is the effect of grammar teaching in English on 5 to 16 year olds' accuracy and quality in written composition?

Research question for in-depth review

What is the effect of teaching *sentence combining* in English on 5 to 16 year olds' accuracy and quality in written composition?

Scope of the review

We have mapped the field of research on the effects of grammar teaching on writing in English speaking countries for pupils aged between 5 and 16 and undertaken an in-depth review of one aspect of the field: the effect of teaching sentence combining on the quality and accuracy of 5–16 year olds' written composition.

For the mapping stage, at least, we looked at empirical research published between 1900 and the present. We limited the review to the teaching of English grammar in schools where English is being taught as a first language (not foreign or second or additional language) in English-speaking countries. We have included research with pupils aged between 5 and 16 and in full-time education. We have focused on the effects of teaching grammar on writing and excluded studies that focus on any effects on reading, or on language acquisition or oracy.

2. METHODS USED IN THE REVIEW

2.1 User involvement

2.1.1 Approach and rationale

The English Review Group involved teachers, school governors, teacher trainers and advisory teachers on its Advisory Group, which commented on and supported the review at each stage. In addition, the results of the review are disseminated more widely through the user summaries, press releases and a journal article, and in seminars (e.g. the ESRC Reconceptualising Writing 5–16 seminar, in which Andrews is a participant), the annual DfES Research Conference and EPPI-Centre dissemination events, etc. Teachers are a principal audience for the results, as are policy-makers (e.g. QCA).

2.1.2 Methods used

User summaries of the review were written by teachers, teacher educators, students, governors and policy-makers. Representatives from each of these constituencies (except students) contributed to the direction and design of the review through the English Review Group's advisory steering committee. Following a meeting with the Teacher Training Agency and PGCE students in June 2003, PGCE tutors and students were involved in a pilot project to write summaries of the present research review (and previous reviews) and to prepare sample lessons arising from the research findings. In addition, a pupil from a secondary school wrote a pupil summary of the final review.

2.2 Identifying and describing studies

2.2.1 Defining relevant studies: inclusion and exclusion criteria

The systematic map included in the systematic review already undertaken by the review team (Andrews *et al*., 2004) was updated for this review. This was done by replication of the methods used for the previous systematic map, for the period May 2003 to April 2004.

For a paper to be included in the systematic map, it had to be a study looking at the effect of grammar teaching in English on 5 to 16 year olds' accuracy and quality in written composition. As the focus of the study is on the *effects* of grammar teaching, papers using methods to identify any such effects were required. This implies the following study types, classified according to the EPPI-Centre taxonomy of study type contained in its core keywording strategy (EPPI-Centre 2002a):

B: Exploration of relationships
C: Evaluation (naturally occurring or researcher-manipulated)
E: Review (systematic or other review) containing at least one study exploring relationships or one evaluation

The full inclusion/exclusion criteria for the review are contained in Appendix 2.1.

2.2.2 Identification of potential studies: search strategy

Reports were identified from the following sources:

- Searching of electronic bibliographic databases: ERIC (Educational Resources Information Center), PsycINFO, SSCI (Social Science Citation Index).
- Searching of reference lists of systematic and other reviews
- Personal contacts

Keywords for searching included the following:

- composition, writing, written composition
- grammar, syntax, text grammar, sentence combining
- metalinguistics
- knowledge about language (KAL)

Appendix 2.2 contains the full search strategy for ERIC, PsycINFO and SSCI.

Searches of these sources from 1900 to 2003 were undertaken for the previous systematic review (Andrews *et al*., 2004) in May 2003. In order to update the systematic map, the searches were re-run for the period May 2003 to April 2004.

2.2.3 Screening studies: applying inclusion and exclusion criteria

The Review Group set up a database system, using EndNote, for keeping track of, and coding, studies found during the update of the review. Titles and abstracts were imported and entered manually into this database. We applied the inclusion and exclusion criteria successively to (i) titles and abstracts, and (ii) full reports. We obtained full reports for those studies that appeared to meet the criteria or where we had insufficient information to make a decision. We reapplied the inclusion and exclusion criteria to the full reports and excluded those that did not meet these initial criteria.

2.2.4 Characterising included studies

The studies remaining after application of the criteria were keyworded using the EPPI-Centre's core keywording strategy (EPPI-Centre, 2002a) and online database software, EPPI-Reviewer (EPPI-Centre, 2002c). Additional review-specific keywords which are specific to the context of the review were added to those of the EPPI-Centre, with definition of the terms in the glossary. The EPPI-Centre's core keywords and the review-specific keywords are contained in appendices 2.3 and 2.4 respectively.

All the keyworded studies were uploaded from EPPI-Reviewer to the EPPI-Centre's Research Evidence in Education Library (REEL), for others to access via the website.

2.2.5 Identifying and describing studies: quality-assurance process

Application of the inclusion/exclusion criteria to the titles and abstracts of studies retrieved through the updated searches was conducted by one member of the Review Group (CT). A second member of the Review Group (AF) applied the criteria to a 25% random sample (45 out of 171) of the titles and abstracts for

quality-assurance purposes. A member of the EPPI-Centre (DE) applied the criteria to a further random sample of 10 titles and abstracts for external quality-assurance purposes. For screening, at the second stage, all papers were double screened by two members of the team working independently (CT and AF). In addition, a member of the EPPI-Centre screened five randomly selected papers for QA purposes. The keywording was conducted by pairs of Review Group members working first independently and then comparing their decisions before coming to a consensus. Members of the EPPI-Centre also helped in applying criteria and keywording for a sample of studies.

2.3 In-depth review

2.3.1 Moving from broad characterisation (mapping) to in-depth review

The inclusion criterion for the in-depth review focused on selected review-specific keywords in order to identify studies that look at the effects of teaching *sentence combining* on the quality and accuracy of pupils' writing

Inclusion criterion for in-depth review

- Must be a study focusing on the teaching of sentence combining

2.3.2 Detailed description of studies in the in-depth review

Studies identified as meeting the inclusion criterion, were analysed in depth using the EPPI-Centre's detailed data-extraction guidelines (EPPI-Centre, 2002b) together with its online software, EPPI-Reviewer (EPPI-Centre, 2002c). Additional questions specific to the context of the review were added to those of the EPPI-Centre.

2.3.3 Assessing quality of studies and weight of evidence for the review question

Three components were used to help in making explicit the process of apportioning different weights to the findings and conclusions of different studies. Such weights of evidence are based on the following:

A Soundness of studies (internal methodological coherence), based upon the study only
B Potential appropriateness of the research design and analysis used for answering the review question
C Relevance of the study topic focus (from the sample, measures, scenario, or other indicator of the focus of the study) to the review question
D An overall weight taking into account A, B and C

To explicate more fully, EPPI-Centre guidelines were used to gauge the weight of evidence an individual study brought to the review. The methodological quality of each study (A) was reviewed in terms of how well it was executed. In addition, each study was assessed for how much weight of evidence (WoE) it provided for the specific review – in terms of (B) the appropriateness of research design for the review question and (C) the relevance of the study for the review question.

Finally, on the basis of judgements about A, B and C, an overall weight (D) was ascribed to each study. This was done on the basis of an approximate average of the three weights A, B and C, although WoE B was given greater importance. A study could only be given an overall WoE of 'high' if it had at least two 'high' judgements, including 'high' for WoE B, and no 'low' judgements. Similarly a study could only be given an overall WoE 'medium' if it had at least two 'medium' (or 'high') WoE judgements, including WoE B. The weight of evidence assessments were taken into consideration in the narrative synthesis. Only studies assessed as 'medium' or 'high' on overall weight of evidence were included in the synthesis.

2.3.4 Synthesis of evidence

The data were then synthesised to bring together the studies which addressed the review questions and which met the quality criteria relating to appropriateness and methodology. A narrative synthesis was undertaken. It was not appropriate to undertake a meta-analysis because the high quality studies in the in-depth review were not sufficiently statistically heterogeneous. The judgements relating to the overall weight of evidence for the included studies were taken into account in the narrative synthesis.

2.3.5 In-depth review: quality-assurance process

Data extraction (including extraction of quantified outcomes data for the three meta-analyses) and assessment of the weight of evidence brought by the study to address the review question were conducted by pairs of Review Group members, working first independently and then comparing their decisions before coming to a consensus. Members of the EPPI-Centre also helped in the data extraction and quality appraisal of a sample of studies.

3. IDENTIFYING AND DESCRIBING STUDIES: RESULTS

3.1 Studies included from searching and screening

Table 3.1 gives the origin of all papers found and those subsequently included in the systematic map. Table 3.2 describes the identification of single studies or reviews that were reported in more than one paper. Figure 3.1 illustrates the process of filtering papers from searching to mapping and finally to synthesis.

Table 3.1: Origin of included papers

	Found May 2003	Found update April 2004	Total papers found	Papers included in map
ERIC	2,557	40	2,597	39
PsycINFO	1,844	115	1,959	2
SSCI	119	16	135	4
Citation	43	1	44	14
Contact	3	3	6	5
Total	***4,566***	***175***	***4,741***	***64***

Papers found on ERIC, PsycINFO and SSCI were imported and de-duplicated hierarchically into the review database. This is reflected in the higher proportion of papers shown as retrieved from ERIC in the original search and included in the map. Interestingly, when the search was updated in April 2004, fewer records were retrieved from ERIC than from PsycINFO. It was established that a new ERIC model for acquiring education literature was being developed and that no new materials had been accepted for the database since December 2003.

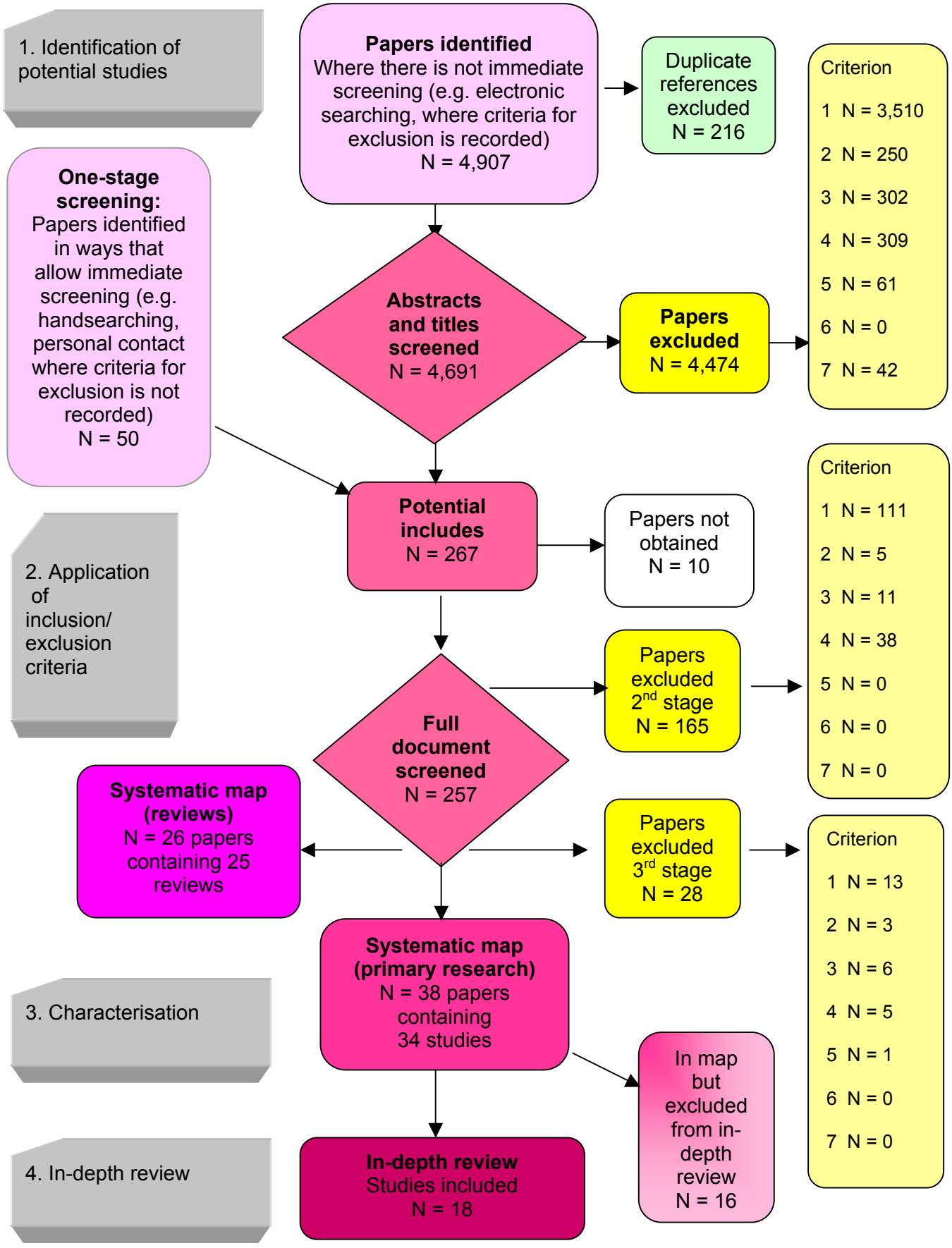
An unusually large number of studies was identified from handsearching the bibliographies of the included systematic and non-systematic reviews, as reflected in the proportionately high number of citations included in the map. Any potentially relevant studies identified through handsearching the reviews were sent for and then screened using the inclusion/exclusion criteria. Any studies that met our inclusion criteria were keyworded and included in the descriptive map (n = 14). A further four studies were identified through expert contact.

Table 3.2: Type of research and number of studies reported by included papers

Research type	Number of papers	Number of reviews or studies
Reviews	26	25
Primary research	38	34

The screening process identified 64 papers that met the inclusion criteria. Table 3.2 shows that 26 papers report reviews and 38 report primary research. One review is reported in two formats: as the full review published in a book and as a summary in a journal article (Hillocks, 1984, 1986). In addition, eight papers (Calkins, 1979, 1980; Combs, 1976, 1977; Elley *et al*., 1975, 1979; Miller and Ney, 1967, 1968) report four studies. The balance of the map therefore describes 25 reviews and 34 studies.

Figure 3.1: Filtering of papers from searching to map to synthesis



3.2 Characteristics of the included studies (systematic map)

Figure 3.2: Publication dates of reviews and studies
(Reviews: N = 25, mutually exclusive)
(Primary studies: N = 34, mutually exclusive)

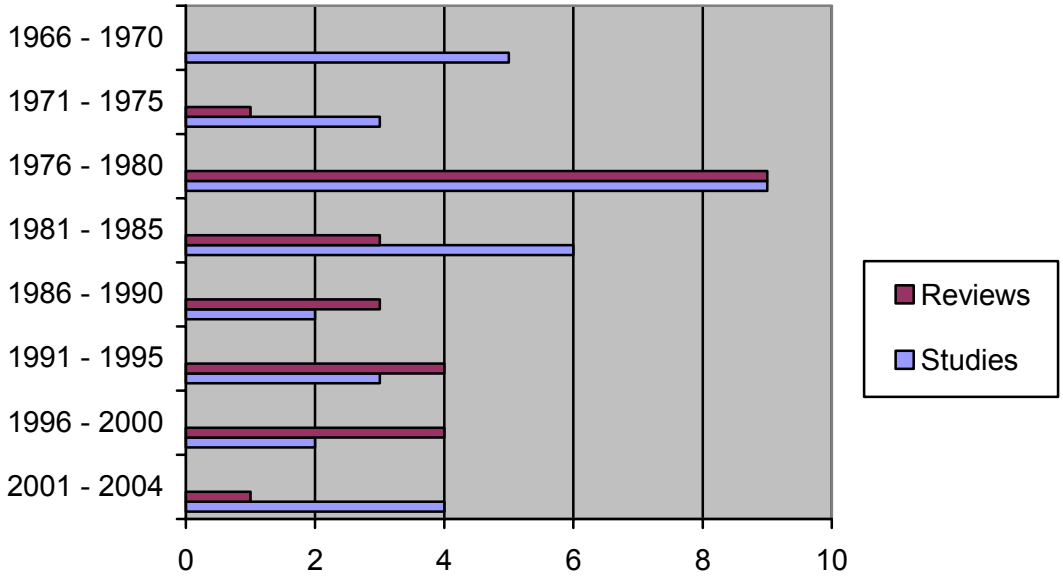


Figure 3.2 defines the publication dates[1] of the included reviews and studies. There was nothing identified that was published before 1966, probably as a result of the electronic searching approach and the particular nature of the review question. However, earlier studies are addressed in the background to the review.

It is interesting to note that 36% (n = 9) of the included reviews were conducted in the five-year period between 1976 and 1980. This contrasts with the ten-year period between 1981 and 1990 when only 24% (n = 6) of included reviews were carried out. However, the proportion starts to rise again between 1991 and 2004, during which a further 36% of included reviews were undertaken. Only one review included in the map was conducted before 1976.

Just under half (n = 15) the primary studies included in the map were conducted in the ten-year period between 1976 and 1985. A further eight of the included studies (24%) were carried out before 1976, with over half (n = 5) of these eight being conducted in the five-year period between 1966 and 1970. A possible explanation for this pattern of publication is the interest in Hunt's theoretical work on T-units and the Subordinate Clause Index (S-C-I) in the

[1] For the purpose of Figure 3.2, 'publication date' is defined as the date that the review or study entered the public domain. As described later in this chapter, a large proportion of the included reviews and studies are in the form of research reports that are unpublished in the sense that they are available only in online databases, such as the ERIC, rather than as journal articles, books, book sections or other conventional media of publication.

1960s (Hunt, 1966). Many of the primary studies used Hunt's S-C-I as an outcome measure.

Only 20% (n = 7) of studies included in the map were conducted in the 15 years from 1986 to 2000. Interestingly, the proportion starts to rise again between 2001 and 2004 with 12% (n = 4) of included studies being carried out in this four-year period. Two of the four studies published in this period are UK publications identified in the update (Green and Sutton, 2003; Holdich *et al*., 2004).

Further characteristics of included reviews

Table 3.3: Type of review (N = 25, mutually exclusive)

Type of review	Number of reviews
Systematic	2
Non-systematic	23
Total	***25***

Almost all the reviews included in the map are non-systematic. Only two (Hillocks, 1984, 1986; Asher, 1990) are systematic reviews.

Table 3.4: Country of origin in which the studies were carried out (N = 25, mutually exclusive)

Country	Number of reviews
USA	16
UK	3
Canada	3
Japan*	1
New Zealand	1
Not stated	1
Total	***25***

*Undertaken by UK academic in Japanese university (Tomlinson, 1994) commenting on the scene in the UK – thus included

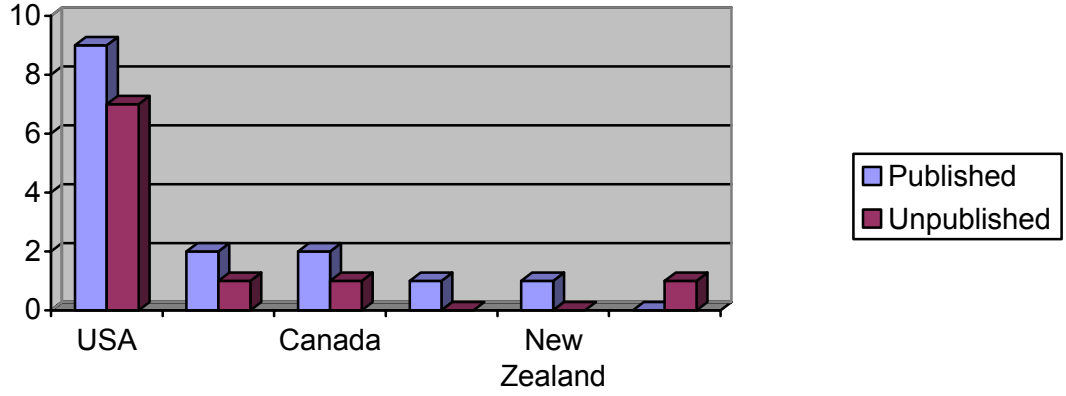
More than half the reviews (63%) were conducted in the USA. One review in eight (n = 3) was conducted in the UK and the same number originated from Canada.

Table 3.5: Publication status (N = 25, mutually exclusive)

Status	Number of reviews
Published	15
Unpublished	10
Total	***25***

Although the majority (60%) of reviews are published, Table 3.5 shows that a high proportion (40%), are in the form of unpublished research reports.

Figure 3.3: Publication status by country of origin (N = 25, mutually exclusive)



The cross-tabulation in Figure 3.3 shows that the status of reviews conducted in the USA is split fairly equally between published (n = 9), and unpublished (n = 7). Of those conducted in the UK and Canada, two of the three studies included for each country are published. Each of the studies conducted in Japan and New Zealand are published. The country of origin of the remaining study (unpublished) is not stated.

Table 3.6: Type of grammar teaching (N = 25, mutually exclusive)

Type of grammar	Number of reviews
Sentence-level	24
Text-level	1

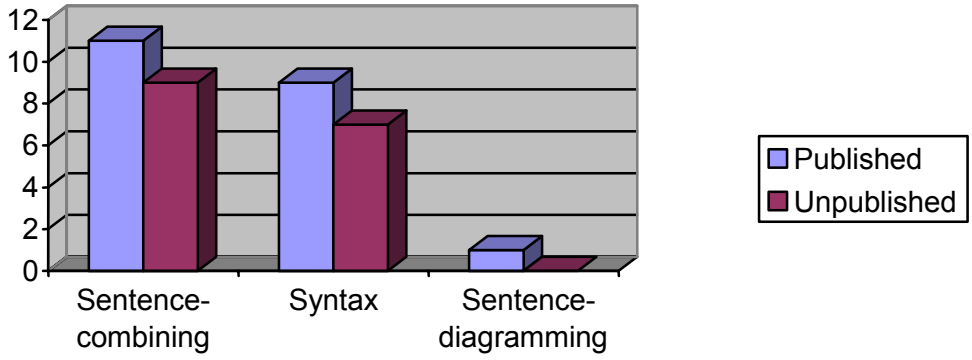
In Table 3.6, we see that 24 of the 25 of the reviews included in the map report on the teaching of sentence-level grammar. Only one review (Seidenberg, 1989) reports on text-level grammar teaching.

Table 3.7: Focus of sentence-level reviews (N = 24, not mutually exclusive)

Focus	Number of reviews
Sentence combining	20
Syntax	16
Sentence-diagramming	1

Of the 24 reviews that report on sentence-level grammar teaching, 20 focused (not exclusively) on sentence combining and 16 focused on syntax (again not exclusively). Twelve reviews report on both sentence combining and syntax, and one review reports on sentence combining, syntax and sentence-diagramming.

Figure 3.4: Reviews of sentence-level teaching by publication status (N = 24, not mutually exclusive)



The cross-tabulation in Figure 3.4 defines the focus of the reviews on sentence-level grammar teaching by publication status. Over half (n = 11) of the 20 reviews on sentence combining are published. Similarly, of the 16 reviews on syntax, nine are published and seven are unpublished. One published review also reports on sentence diagramming. The in-depth review focuses on the effect of teaching sentence combining. Therefore the conclusions of the 20 reviews that include a focus on sentence combining are presented in Table 3.8.

Table 3.8: Summary of conclusions of the two systematic and 18 non-systematic sentence-combining reviews

Author, date		**review**	**f teaching grammar on writing**
Systematic reviews			
Asher, 1990	Not stated	27	'From these initial results of a large body of experimental literature in writing instruction, it seems possible to conclude that sentence combining works best at the junior high school level, but its effectiveness drops both before and after that level, an interesting non-linear relationship' (p 152).
Hillocks, 1984, 1986	'The study of parts of speech, and sentences'	5 (grammar) 5 (sentence combining)	Grammar: '…every other focus of instruction examined in this review is stronger' (1984, p 160). 'Sentence combining': 'on the average, ..(is) ..more than twice as effective as free writing as a means of enhancing the quality of student writing' (1984, p 161).
Non-systematic reviews			
Abrahamson, 1977	Not stated	8 evaluative abstracts, but 7 empirical studies	'…the study concludes that traditional grammar instruction does not help students improve their writing ability appreciably, that such instruction, in fact, may hinder the development of students as writers, and that sentence-combining instruction should be incorporated into both elementary and secondary language arts programs' (p 1).
Amiran and Mann, 1982	Not stated	5	'…study of traditional grammar, per se, is not recommended as a primary activity' (p 51). '… activities such as sentence combining and the study of transformational generative grammar using non-traditional work exercises are most successful' (p 51).
Crowhurst, 1980	Not stated	8 interventions but 7 relevant studies	'First, neither T-unit length nor clause length is a good predictor of writing quality. Second, although sentence-combining studies sometimes seem to improve writing quality, the improvement is probably due to factors other than increases in T-unit and clause length' (p 2).

Author, date	Definition of 'grammar'	Number of studies in review	Conclusion – effect of teaching grammar on writing (sentence combining)
Gann, 1984	Not stated	2	'…grammar instruction almost certainly does not contribute significantly to improvement in written English' (p 49).
Hudson, 2000	Not stated	13 reviews, 28 further separate studies	Hudson declares at the end that 'the idea that grammar teaching improves children's writing skills is much better supported by the available research than is commonly supposed' but his review shows that traditional grammar teaching is ineffective, on the whole, whereas sentence combining is effective.
Kolln, 1996	Not stated	4	No conclusions
Lawlor, 1980	Detailed definitions of transformational generative grammar theory and sentence combining (pp 1–3)	12	'there is some evidence that sentence combining can lead to an overall improvement in writing quality' (p 62). 'Sentence combining is not dependent upon a formal knowledge of grammar' (p 63). 'One of the most significant conclusions that can be drawn from the research is that sentence combining can be an effective strategy in nearly every grade level across the academic spectrum' (p 64),
Matzen *et al.*, 1995	Not stated	5	No conclusions
Ney, 1980	Not stated	8	'Sentence combining should be an important part of the elementary school curriculum' (p175).
Phillips, 1996	Not stated	6	Author draws no conclusions: 'a few researchers question the use of the T-unit and forced choice ratings in various sentence combining studies but most researchers praise sentence-combining as an effective means of improving written composition' (p 2).
QCA, 1998a	Not stated 'Grammar teaching is a complex issue'.	10	'Discrete teaching of parts of speech and parsing in decontextualised form is not a particularly effective activity…There is no evidence that knowledge acquired in this way transfers into writing competence.' 'Transformational-generative grammar…has little to offer…' 'There is evidence from studies of writing development that experience of the syntactic demands of different types of tasks is a key factor in pupils' written performance and development' (p 55).
Sternglass, 1979	Not stated	2	No conclusions
Stewart, 1979	Not stated	7	No conclusions
Stotsky, 1975	Not stated	7	'Writing behavior appeared to be significantly altered in the direction of more mature syntactic structures' (p 54). 'Results from the experimental writing programs have confirmed the intuition of many educators of the value in helping students acquire explicit facility with syntactic structures … insofar as it affects their writing skills' (p 63).
Ulin and Schlerman, 1978	Not stated	4	'…these studies have shown that TG [transformational grammar] is no more effective than traditional grammar in improving composition…' (p 65). 'In summary, evidence suggests that sentence-combining exercises might improve composition in ways that grammar has long been alleged to do, particularly in the areas of sentence structure, usage, ideas and style' (p 65).

Author, date	Definition of 'grammar'	Number of studies in review	Conclusion – effect of teaching grammar on writing (sentence combining)
Walsh, 1991	'…the system of word structures and word arrangements of the language' (p 3)	13	'…it does not follow that knowledge of grammar will make one a better writer' (p 7).
White and Karl, 1980	Refers to 'transformational grammar theory from which sentence-combining is derived' (p 227) but no definitions are given.	7	'Sentence-combining studies have consistently recorded significant gains in writing improvement at all levels (elementary, secondary and college) ' (p 226).
Wyse, 2001	It cites five definitions of grammar by Hartwell (1985) and refers to structural, transformational, generative, functional grammars, etc. 'This paper focuses attention on some of the empirical evidence in relation to Traditional School Grammar (TSG), transformational grammar, and sentence-combining' (p 412).	15 reviews, 12 individual studies	'The findings from international research clearly indicate that the teaching of grammar (using a range of models) has negligible positive effects on improving secondary pupils' writing' (p 422). 'The one area where research has indicated that there may be some specific benefit for syntactic maturity is in sentence-combining' (p 423).

The conclusions in these reviews are used to contextualise our results in the discussion section of Chapter 5.

Further characteristics of included primary studies

Of the 34 primary studies included in the map, almost all are study type C (i.e. evaluations). We searched and screened for study type B (exploration of relationships) but found only one that met our inclusion criteria.

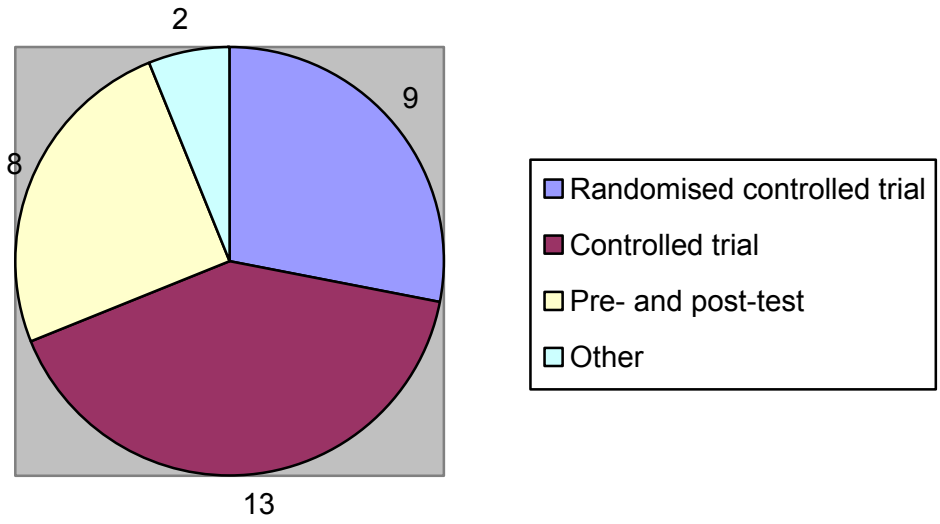
Table 3.9: Type of study (N = 34, mutually exclusive)

Type of study	Number of studies
Researcher-manipulated evaluation	32
Naturally-occurring evaluation	1

Exploration of relationships	1
Total	***34***

Table 3.9 shows that almost all the evaluations of primary research included in the map were researcher-manipulated. Only one evaluation was found to be of a naturally-occurring intervention.

Figure 3.5: Type of researcher-manipulated evaluation (N = 32, mutually exclusive)



Of the 32 researcher-manipulated evaluations, nine report randomised controlled trials, 13 report controlled trials, eight report pre-and post-tests, and two report other types of evaluation.

Table 3.10: Country of origin (N = 34, mutually exclusive)

Country	Number of studies
USA	28
Canada	3
UK	2
New Zealand	1
Total	***34***

Table 3.10 shows that 82% of primary studies (n = 28) included in the map originated in the USA. Three studies were conducted in Canada, two in the UK and one in New Zealand.

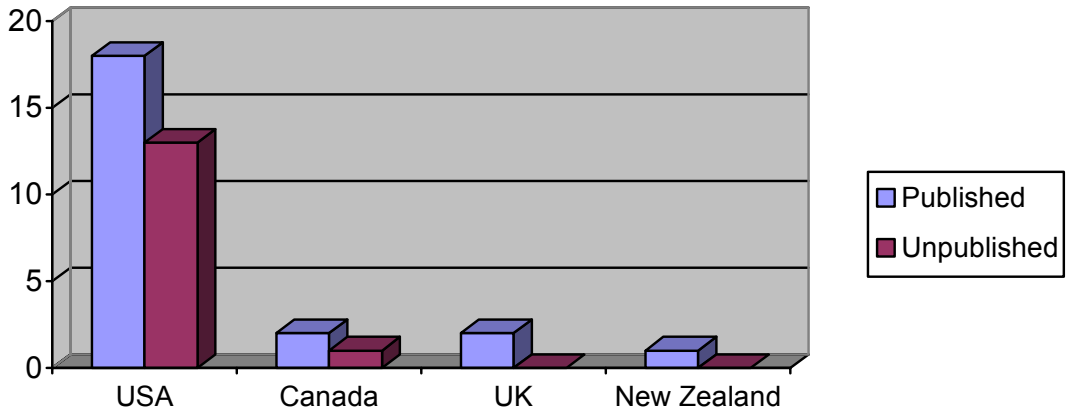
Table 3.11: Publication status (N = 34, mutually exclusive)

Status	Number of studies
Published	20
Unpublished	14

Total	***34***

As in the case of the reviews, the majority of studies (59%) are published, but a high proportion (41%) are unpublished.

Figure 3.6: Publication status by country of origin (N = 34, mutually exclusive)



The cross-tabulation in Figure 3.6 shows that the status of primary studies originating in the US is again split almost equally between published (n = 18), and unpublished (n = 13). Of the three studies conducted in Canada, one is published and one unpublished. The two UK studies are published, as is the remaining New Zealand study.

Table 3.12: Types of Pupils (N = 34, not mutually exclusive)

Educational setting	**Number of studies**
Primary school	13
Secondary school	19
Special needs school	3
Independent school	2
Residential school	1
Age of pupils	
5–10	17
11–16	25
Sex of pupils (mutually exclusive)	
Mixed sex	15
Male only	2
Female only	1
Not stated	16

Table 3.12 describes the educational settings in which the studies were conducted and the age and sex of the pupils involved. Four studies were conducted in more than one educational setting and eight studies involved pupils in both primary and secondary school age groups. In just under half of the

studies, the sex of pupils was not stated. Of the remaining 18 studies, the majority involved pupils of both sexes (n = 15).

Table 3.13: Type of grammar teaching (N = 34, not mutually exclusive)

Type of grammar	Number of studies
Sentence-level	29
Text-level	9

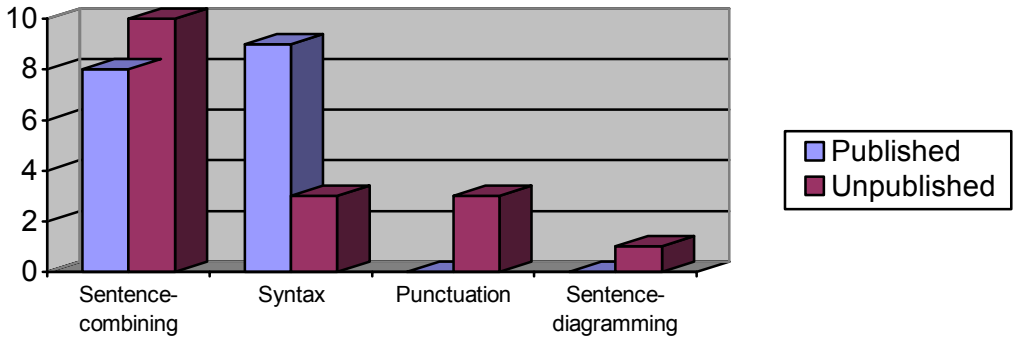
In Table 3.13 we see that 29 studies (85%) focused on the teaching of sentence- level grammar. Nine studies focused on text-level teaching and four involved the teaching of both types of grammar.

Table 3.14: Focus of sentence-level studies (N = 29, not mutually exclusive)

Focus	Number of studies
Sentence combining	18
Syntax	12
Punctuation	3
Sentence-diagramming	1

Of the 29 studies that report on sentence-level grammar teaching, 18 focused on sentence combining and 12 focused on other aspects of syntax. A much smaller proportion focused on punctuation (n = 3), and only one study focused on sentence-diagramming. Three studies investigated the teaching of both sentence combining and syntax. One study focused on sentence combining and punctuation, one on syntax, punctuation and sentence-diagramming, and one on punctuation alone.

Figure 3.7: Studies of sentence-level teaching by publication status



(N = 29, not mutually exclusive)

It is interesting to note from the cross-tabulation in Figure 3.7 that, of the 18 studies focusing on sentence combining, just over half (n =10) are unpublished. The reverse is true of the studies concerned with syntax. In this group, nine are published and three unpublished. The balance of the studies on punctuation (n = 3) and sentence-diagramming (n = 1) are unpublished.

Table 3.15: Focus of text-level studies (N = 9, not mutually exclusive)

Focus	Number of studies
Paragraph composition	5
Text structure	4
Cohesion	2

Of the nine studies that report on text-level grammar teaching, just over half (n = 5) focused on paragraph composition. Four studies focused on text structure and two on cohesion. One study investigated the teaching of both paragraph composition and text structure, and one study investigated both paragraph composition and cohesion.

Figure 3.8: Studies of text-level teaching by publication status
(N = 9, not mutually exclusive)

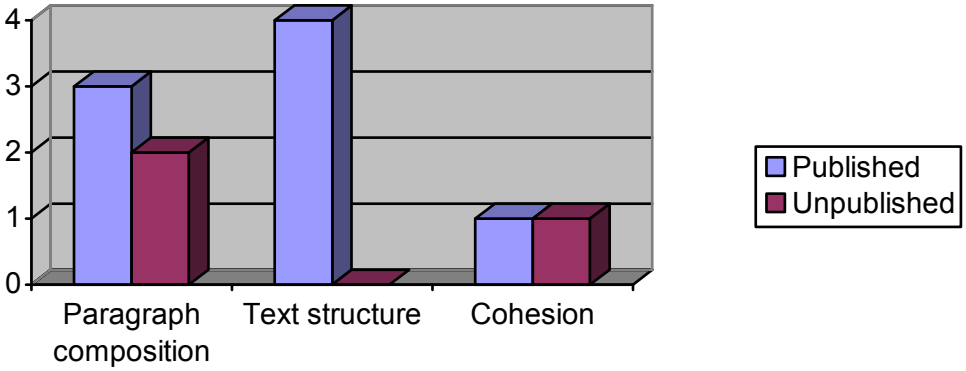


Figure 3.8 shows that, of the five studies concerned with paragraph composition, three were published and two were unpublished. All four studies focusing on text structure were published and, of the two studies that focused on cohesion, one was published and one was unpublished.

3.3 Identifying and describing studies: quality-assurance results

Quality assurance of the two stages of screening papers retrieved from the electronic searches

(i) Screening of titles and abstracts (CT, AF and DE)
There were 171 records in the update library that were screened by CT. A random sample of 45 of these was independently screened by AF. There was full agreement to include or exclude on 44 of these studies. AF included one study that CT excluded. We looked at the study together and agreed to exclude it. DE (from the EPPI-Centre) independently screened a further random sample of 10 titles and abstracts. There was full agreement with CT on 7 of these reports. DE was more inclusive and included three studies that CT excluded. Decisions to exclude were made on the basis of discussion. It was agreed between CT, AF and DE that it was not necessary to take the QA procedures for screening any further as it was felt that the Review Group could have confidence in CT's screening of the updated database.

(ii) Screening of full papers (CT, AF and DE)
The studies identified as being potentially relevant from the screening of the database, or from expert contact, were sent for and independently double screened on the basis of the full papers by CT and AF. Full agreement was established on whether or not to include or exclude; that is, CT and AF agreed on all studies. DE independently double screened five studies. Of the five studies, there was full agreement on four to include or exclude (include one, exclude three). DE also excluded one of the studies that had been included by CT and AF (Asher, 1990). On looking again at this study, DE agreed to include. This was a methodological paper that contained an update of the Hillocks (1984) systematic review. Therefore it was included as a 'review'.

Quality assurance of keywording

Keywording
All of the studies that were included in the update were independently double keyworded by either AF and CT or by AF and DE. Agreement was generally very high on all generic and review-specific keywords. If there was any disagreement, this was resolved through discussion.

4. IN-DEPTH REVIEW: RESULTS

4.1 Selecting studies for the in-depth review

To be included in the in-depth review a study had to report on the teaching of sentence combining. The application of the inclusion criterion described in section 2.3.1 identified 18 studies for in-depth review, as follows:

Table 4.1: Studies included in in-depth review

Author(s), year and title
Combs WE (1976) Further effects of sentence-combining practice on writing ability Combs WE (1977) Sentence-combining practice: do gains in judgements of writing 'quality' persist?
Combs WE, Wilhelmsen K (1979) In-class 'action' research benefits research, teacher and students
Hunt KW, O'Donnell R (1970) An elementary school curriculum to develop better writing skills
MacNeill TB (1982) The effect of sentence-combining practice on the development of reading comprehension and the written syntactic skills of ninth grade students
McAfee D (1981) Effect of sentence combining on fifth grade reading and writing achievement
Mellon J (1969) Transformational sentence combining
Melvin MP (1980) The effects of sentence combining instruction on syntactic maturity, reading achievement and language arts skills achievement
Miller BD, Ney JW (1967) Oral drills and writing improvement in the fourth grade Miller BD, Ney JW (1968) The effect of systematic oral exercises on the writing of fourth-grade students
Ney JW (1976) Sentence combining and reading
Nutter N, Safran SP (1983) Sentence combining and the learning disabled student
O'Hare F (1973) Sentence combining: improving student writing without formal grammar instruction
Pedersen EL (1978) Sentence-combining practice: training that improves student writing
Roberts CM, Boggase BA (1992) Non-intrusive grammar in writing
Rousseau MK (1989) Increasing the use of compound predicates in the written compositions of students with mild learning handicaps
Rousseau MK, Poulson CL (1985) Using sentence-combining to teach the use of adjectives in writing to severely behaviorally disordered students
Saddler B, Graham S (forthcoming) The effects of peer-assisted sentence combining instruction on the writing performance of more and less skilled young writers
Stoddard EP, Renzulli JS (1983) Improving the writing skills of talent pool students
Vitale MR, King FJ, Shontz DW, Huntley GM (1971) Effect of sentence-combining exercises upon several restricted written composition tasks

Two of these studies (Combs 1976, 1977; Miller and Ney 1967, 1968) were reported in linked pairs of papers. One paper was selected as the lead paper

for each study, but data in both papers were drawn on for data-extraction purposes.

4.2 Comparing the studies selected for in-depth review with the total studies in the systematic map

Country of origin

Seventeen of the 18 studies selected for in-depth review originated in the USA. One (MacNeill, 1982) was carried out in Canada. This mirrors the high proportion of USA studies in the systematic map.

Publication status

The proportion of studies that are published is reversed when compared with that of all the studies in the map. Of the studies selected for in-depth review, 39% (n = 7) are published and 61% (n = 11) are unpublished. This compares with 59% of studies in the map that are published and 41% that are unpublished.

Table 4.2: Publication status of studies selected for in-depth review

Publication status	Number of studies	Study
Published	7	Combs (1976, 1977) Combs and Wilhelmsen (1979) Mellon (1969) Miller and Ney (1967, 1968) O'Hare (1973) Stoddard and Renzulli (1983) Vitale (1976)
Unpublished	11	Hunt and O'Donnell (1970) MacNeill (1982) McAfee (1981) Melvin (1980) Ney (1976) Nutter and Safran (1983) Pederson (1978) Roberts and Boggase (1992) Rousseau (1989) Rousseau and Poulson (1985) Saddler and Graham (forthcoming)*

* One study (Saddler and Graham, forthcoming) was in press at the time of writing this report.

Study type

All the studies selected for in-depth review report researcher-manipulated evaluations and the proportions of RCT, CT and pre- and post-test designs are similar to those of all studies included in the map.

Table 4.3: Type of study selected for in-depth review

Study type	Number of studies	Study
Randomised controlled trial	5	McAfee (1981) O'Hare (1973) Saddler and Graham (forthcoming) Stoddard and Renzulli (1983) Vitale (1976)
Controlled trial	9	Combs (1976, 1977) Combs and Wilhelmsen (1979) Hunt and O'Donnell (1970) MacNeill (1982) Mellon (1969) Melvin (1980) Miller and Ney (1967, 1968) Nutter and Safran (1983) Pederson (1978)
Pre- and post-test	4	Ney (1976) Roberts and Boggase (1992) Rousseau (1989) Rousseau and Poulson (1985)

4.3 Further details of studies included in the in-depth review

Appendix 4.1 provides summary tables of the 18 studies included in the in-depth review. These tables are based on the information gathered and judgements reached in the data extraction of the studies.

4.4 Synthesis of evidence

4.4.1 Weight of evidence

Table 4.4 sets out the ranking of the studies in the in-depth study in terms of weight of evidence.

Table 4.4: Ranking of weight of evidence of studies in the in-depth study

Studies	Weight of evidence A (trustworthiness in relation to study questions)	Weight of evidence B (appropriateness of research design and analysis)	Weight of evidence C (relevance of focus of study to review)	Weight of evidence D (overall weight of evidence)
O'Hare (1973)	High	High	High	High
Saddler and Graham (forthcoming)	High to medium	High	Medium	High to medium
Combs (1976, 1977)	High	Medium	High to medium	Medium to high
Hunt and O'Donnell (1970)	Medium to high	Medium	Medium to high	Medium to high
Combs and Wilhelmsen (1979)	Medium	Medium	High	Medium
Vitale *et al.* (1971)	Medium	High	Medium	Medium
MacNeill (1982)	Medium to low	High	Medium to high	Medium
Stoddard and Renzulli (1983)	Medium	High to medium	Medium	Medium
Mellon (1969)	Medium to high	Medium	Medium	Medium
Miller and Ney (1967, 1968)	Medium	Medium	Medium	Medium
Pedersen (1978)	Medium to low	Medium	Medium	Medium
Melvin (1980)	Low	Medium	Medium	Medium to low
Nutter and Safran (1983)	Low	Medium	Medium	Medium to low
Rousseau and Poulson (1985)	Medium	Low	Medium	Medium to low
McAfee (1981)	Low to medium	High to medium	Low to medium	Low to medium
Ney (1976)	Medium to low	Medium to low	Low	Low
Roberts and Boggase (1992)	Low	Low	Medium	Low
Rousseau (1989)	Medium	Low	Low	Low

For the purposes of the narrative synthesis, we only included those studies ranked of medium weight of evidence overall or above. This does not mean to say that those studies ranked medium to low and below are not worthy studies; it simply means that, for the purposes of answering the specific research question in the present systematic review, those ranked medium and above provide the best evidence.

4.4.2. The four best studies for the purposes of answering our research question

The four best studies for the purposes of answering our research question are O'Hare (1973), Saddler and Graham (forthcoming), Combs (1976, 1977), Hunt and O'Donnell (1970). Table 4.5 gives a definition of what sentence combining means in each of these four studies. It also cross-references the sentence-combining techniques outlined in the background section with what was actually done in these key research studies.

Table 4.5: Key research studies (sentence-combining definitions and techniques)

Author(s), year	Definition	Technique(s)
O'Hare (1973)	This study's system of sentence-combining practice is described as 'practice with intensive sentence manipulation that involved multiple embedding of kernels supplied in advance' (p 34).	Embedding techniques used in this study were a replication of those used by Mellon (1969): 'the experimental group was required to write out sentences virtually identical to those written out by Mellon's experimental group' (p 37). In Mellon's research, 'the student was given a set of kernel sentences plus directions for combining these sentences into a single complex statement' (Mellon, 1969, p 32).
Saddler and Graham (forthcoming)	'The intervention employed in the present study directly taught students how to construct more complex and sophisticated sentences by combining two or more basic (i.e., 'kernel') sentences into a single sentence (Ney, 1981; Strong, 1976). This instructional method, referred to as sentence combining, is not only designed to teach students how to craft more syntactically complex sentences, but to produce better sentences, ones that more closely convey the writer's message' (p 6).	Compounding and embedding 'Instruction was broken down into five units … the first unit focused on combining smaller related sentences into a compound sentence' (p 13). 'The next unit involved embedding an adjective or adverb from one sentence into another … The third and fourth units concentrated on creating complex sentences by embedding an adverbial and adjectival clause, respectively, from one sentence into the other… The final unit extended the embedding skills taught in units 2–4 by teaching students to make multiple embeddings' (p 14).
Combs (1976, 1977)	This study is a replication of the Mellon (1969) and O'Hare (1973) studies. Sentence combining is not precisely defined; however, a description of Mellon's exercises is given: 'The exercises followed a standard format; each was a set of kernel sentences plus directions for combining the kernels into a single complex statement to be written out by the student' (p 137).	This study replicates those of Mellon and O'Hare and necessarily uses their techniques. However, these techniques are not explicitly stated.

Hunt and O'Donnell (1970)	No definition is given, although reference is made to the work of Mellon (1969).	The study 'required that several sentences … be embedded in one another' (p 4).

O'Hare's (1973) study arguably represents the apex of studies on the effect of sentence combining on written composition. Its aim was 'to test whether sentence-combining practice that was in no way dependent on the students' formal knowledge of transformational grammar would increase the normal rate of growth of syntactic maturity in the students' free writing in an experiment at the seventh grade level over a period of eight months' (p 35). Within a total sample of 83, students were randomly assigned to two experimental and two control classes, thus creating a randomised controlled trial. Pre- and post-tests were undertaken on three kinds of writing sample: narration, description and exposition; and six factors of syntactic maturity were employed: words per T-unit, clauses per T-unit, words per clause, noun clauses per 100 T-units, adverb clauses per 100 T-units and adjective clauses per 100 T-units. This particular study is comprehensive, with high degrees of validity and reliability.

Results from the study show that not only did the experimental group experience highly significant growth, but that its performance exceeded that of the control group on all six measures of syntactic maturity. Indeed, eighth graders were writing at the same syntactic maturity level as twelfth graders on five of the six measures. In terms of writing quality, as judged and agreed by a team of eight evaluators, the experimental group also exceeded the performance of the control group, particularly in narrative and descriptive composition. The author of the study concludes that 'teachers of writing surely ought to spend more time teaching students to be better manipulators of syntax. Intensive experience with sentence-combining should help enlarge a young writer's repertoire of syntactic alternatives and to supply him [sic] with practical options during the writing process' (p 76).

Saddler and Graham's study (forthcoming) suggests that sentence combining is not a practice nor a topic for research confined to the 1960s to 1980s. The aim of their study was to examine the effectiveness of sentence-combining instruction, coupled with peer instruction, 'for improving a basic foundation writing skill, sentence construction' (p 4). The assumption behind this particular study is that facility in generating sentences should make available more cognitive resources for other aspects of composition. Using a sample of 44 9–11 year old pupils (the mean participating age was 9 years, 3 months), the authors used sentence combining or grammar interventions to pupils in pairs in laboratory-like conditions. The study type was that of an individualised randomised controlled trial with stratified randomisation; baseline equivalence was used to eliminate chance bias. Although the reliability of the study was high, validity would seem to be less strong than in O'Hare or other studies. However, the results are clear: 'sentence-combining instruction was effective in improving the sentence-combining skills' (p 29) and has a positive impact on writing quality, not only in first versions of writing but also in subsequent revisions. The effect of sentence combining was seen to be stronger in the development of syntactic maturity than in the improvement in writing quality. The writers conclude that 'findings from the current study replicate and extend previous research by showing that a peer-assisted sentence-combining treatment can improve the sentence construction skills of more and less skilled young writers…and that such instruction can promote young students' use of sentence-combining skills as they revise' (p 37).

O'Hare (1973) and Saddler and Graham (forthcoming) were thought to have the highest weight of evidence in relation to the research question set by the review. The following studies provided medium weight of evidence overall. Of these, the studies by Combs (1976, 1977) and Hunt and O'Donnell (1970) were afforded medium to high weight of evidence.

Combs' studies replicated aspects of earlier studies by Mellon (1969) and O'Hare (1973) with a sample of 100 seventh grade students. The design of the study 'included two intact experimental classrooms and two intact control classrooms selected from a suburban Minneapolis junior high school and followed the pre-test control group design…excepting the random selection of the student population and the inclusion of a delayed post-test'. In effect, this was a clustered controlled trial. Narrative and descriptive modes of writing were used to provide writing samples and seven teacher-raters were used to gauge the quality of matched pairs of writing from the control and experimental groups. The study was relatively well conducted in terms of validity and reliability, and its results show that using words per T-unit and words per clause – the two most discriminating measures in terms of syntactic maturity – revealed that students made a grade leap of + 2, as opposed to Mellon's (+ 1) and O'Hare's (+ 5). Although the experimental period was shorter than in the study by O'Hare, it is suggested that the delayed post-test in Combs' study indicates a more reliable measure of sustained syntactic progress. With both syntactic maturity scores and overall quality of writing improved, the author concludes that 'sentence-combining practice seemed to affect more than syntactic gains, indeed, gains that were incorporated in what teacher-raters consider improved quality of writing' (p 321). The correlation between syntactic maturity gains and overall writing quality is not clearly described, however. Caution is required in interpreting the results of these papers by Combs, as the trial sample was not randomised.

Hunt and O'Donnell's (1970) study, which used a sample of 335 students, was again a clustered trial without randomisation. Its aim was to examine the impact of sentence combining on the writing of fourth grade students, specifically with 194 black and 141 white students. Again, the measures used to gauge syntactic maturity were words per T-unit, clauses per T-unit and words per clause. As in Combs' studies, gains were two grade levels for the experimental groups, with particular gains in syntactic maturity for black students; but there was no delayed post-test, so gains might have been short-term. In general, this is a study with high validity and reliability, with a relatively large sample, but constrained by the fact that the pre-test did not include a writing sample and the fact that there was no delayed post-test. Finally, the clustered trial nature of the study, without randomisation, means that we cannot be sure that other factors not mentioned in the study bore some influence on the results.

If we look at the effect sizes of the four studies mentioned so far – the four that provide the best evidence in answering our research question – we can see that, with regard to the outcome measure of words per T-unit (which is regarded by these authors as the best measure of syntactic maturity), O'Hare (1973) finds a very large positive effect for the intervention of sentence combining on writing accuracy and quality (effect size = 2.4, CI 1.81 to 2.94). In studies by Combs (1976, 1977), this effect (post-test effect 1.09, CI 0.66 to 1.50) is confirmed, but found to lessen somewhat as measured by delayed post-test

(effect 0.68, CI 0.27 to 1.07). All three of these results are statistically significant.

Effect sizes from the four studies with greatest overall weight of evidence

Table 4.6: Effect sizes: O'Hare (1973)

Autor, date: O'Hare, 1973				
Outcome	**Hedges 'g'**	**CI Upper**	**CI Lower**	**Effect size***
Words/T-unit	2.4	2.94	1.81	Very large positive*
Clauses/T-unit	1.94	2.44	1.4	Very large positive*
Words/clause	1.70	2.18	1.81	Large positive*
Noun clauses/ 100T-units	0.75	1.19	0.3	Fairly large positive*
Adverb clauses/100T-units	1.41	1.88	0.91	Very large positive*
Adjective clauses/100T-units	1.86	2.37	1.37	Very large positive*

* denotes statistical significance

Table 4.7: Effect sizes: Combs (1976, 1977)

Author, date: Combs, 1976 and 1977				
Outcome	**Hedges 'g'**	**CI Upper**	**CI Lower**	**Effect size***
W/T-U post-test	1.09	1.50	0.66	Very large, positive*
W/T-U delayed post-test	0.68	1.07	0.27	Fairly large, positive*
W/C post-test	0.56	0.96	0.16	Moderate, positive*
W/C delayed post-test	0.45	0.84	0.05	Moderate, positive*

* denotes statistical significance

Hunt and O'Donnell (1970) use four measures: a global score for free writing, words per clause, clause per T-unit and words per T-unit. The last three outcomes are comparable with O'Hare (1973) and Combs (1976, 1977). The words per clause measure for O'Hare (1973) shows a large positive effect with statistical significance; Combs (1976, 1977), as in the measures of words per T-Unit, shows a more moderate positive effect. Hunt and O'Donnell (1970) show a borderline positive effect on this particular measure. On the words per T-unit measure, their result is the demonstration of a moderate positive effect, similar to that of Combs (1976, 1977).

Table 4.8: Effect sizes: Hunt and O'Donnell (1970)

Author, date: Hunt and O'Donnell, 1970				
Outcome	**Hedges 'g'**	**CI Upper**	**CI Lower**	**Effect size***
Free writing total scores (all students covaried by pre-test)	0.911	1.135	0.683	Large and positive*
Free writing words/clause (all students covaried by pre-test)	0.0024	0.217	–0.212	Borderline positive
Free writing clause/T-unit (all students covaried by pre-test)	1.216	1.447	0.979	Large and positive*
Free writing words/T-unit	0.339	0.5552	0.1226	Moderate and positive*

* denotes statistical significance

Saddler and Graham (forthcoming) use a different measure for syntactic maturity from the other three highly rated studies: a sentence-combining test, which shows a large, positive effect size with statistical significance.

Table 4.9: Effect sizes: Saddler and Graham (forthcoming)

Author, date: Saddler and Graham, 2004				
Outcome	**Hedges 'g'**	**CI Upper**	**CI Lower**	**Effect size***
TOWL-3 sentence-combining test	0.929	1.543	0.281	Large and positive*
Story quality	0.423	1.02	–0.189	Moderate and positive
Story length	–0.216	0.387	–0.812	Small and negative

* denotes statistical significance

4.4.3 Publication bias

It is possible that our results are affected by publication bias. Although we included research reports and conference papers (studies that are unpublished but in the public domain), we excluded unpublished PhD theses. We were unable to investigate the possibility of the presence of publication bias in the review through the use of a funnel plot because we only had four high quality trials with effect sizes in our in-depth review and this is too few for a funnel plot. However, it seems probable that our results are affected by publication bias, given that the two 'unpublished' studies (Saddler and Graham, forthcoming; Hunt and O'Donnell, 1970) have smaller effect sizes than the two published studies (Combs, 1976 and 1977; O'Hare, 1973). Therefore the findings and conclusions of our review should be treated with caution.

4.4.4 Other studies providing lower weight of evidence

The remaining five studies that were rated medium for one reason or another provide somewhat less weighty evidence for the efficacy of sentence combining on writing accuracy and quality.

Combs and Wilhelmsen's (1979) experimental group students in eighth grade showed gains in syntactic maturity, 'particularly in their argumentative papers' (p 269) which is interesting when we compare the results to those of writers above who found that it was in narrative and descriptive modes that the greatest gains were made. Their assumption that the argumentative mode 'lends itself to more complex syntax' cannot be taken too seriously: if we compare a Henry James narrative sentence to that of a rhetorically repetitive argument, it becomes hard to generalize about syntactic maturity in one mode rather than another. The strength of this study, however, is in its contextual 'action-research' location within a classroom, rather than being conducted in a laboratory-like environment. The authors conclude that, because the control group was taught formal grammar, 'if one wants students to do well on grammar tests, teach them grammar. But do not expect it to have measurable influences on their writing maturity. However, if one wants the grammatical structures to show in students' free writing, then present them with sentence-combining strategies in systematic, extended, and creative lessons' (p270).

Of the remaining studies, Mellon (1969) in a study of 247 seventh grade students, found 'the experimental group experienced significant pre-post growth on all twelve factors' with the control group's advances being 'so slight as to be virtually indiscernible' (p 74). Miller and Ney (1967, 1968) found similar results with a fourth grade class; and Pedersen (1978), with 113 seventh grade students, found that such students 'trained in sentence combining scored significantly higher than control subjects in achieving and sustaining growth in syntactic fluency' (p 7) as well as in conceptualisation and expression. All these three studies used a clustered controlled trial study type without randomisation.

The only paper not to show gains in syntactic maturity of those that were rated of medium weight of evidence or above in the in-depth review was that by MacNeill (1982). This randomised controlled trial with a sample of 154 (with 11 lost due to 'experimental mortality') was designed to investigate the effect of the O'Hare 'Sentencecraft' programme on the written syntactic skills of ninth graders, using argumentative compositions as the data source. Although six major indices of syntactic maturity were used, including words per T-unit and words per clause, no significant mean increases were found in the writing of the experimental group students using this particular programme. And, although insufficient data on writing outcomes were presented to reviewers to calculate effect sizes, the randomised controlled trial was an appropriate design to use in trying to answer the research question. It is a pity that the reporting of the data did not provide more evidence, as this study is the only one among those with greatest weight of evidence to run counter to the general trend.

4.5 In-depth review: quality-assurance results

All the studies in the in-depth review were independently double data extracted by two reviewers who then compared their data extractions and resolved any differences. This procedure was followed for extraction of all data (including outcomes data for calculating effect sizes) and for quality appraising the studies and applying the weight of evidence judgements. The EPPI-Centre link people also double data extracted two studies for quality-assurance purposes. The results of the double data extractions were that agreement was high. Occasional disagreements about weight of evidence were discussed and resolved.

4.6 Nature of actual involvement of users in the review and its impact

As with other systematic reviews by the English Review Group, the steering committee played a significant role in suggesting the focus of the study, in reading a draft of the protocol, and in reading a draft of the final report. The group consists of primary and secondary teachers, parent governors, a Chair of a governing body, English advisory teachers, researchers, health studies experts, teacher educators and policy-makers.

Independent peer review at protocol stage and at draft stage of the final report provided another dimension of involvement and critique, as did discussion of the emerging findings and methodology at an Economic and Social Research Council seminar series on ‘Reconceptualising Writing 5–16’ which involved teachers, members of the Qualifications and Curriculum Authority, teacher educators and researchers…many of whom will have been parents of school-age children and/or governors of schools.

Finally, initial and provisional findings were discussed at the annual conference of the British Educational Research Association in Manchester, September 2004.

5. FINDINGS AND IMPLICATIONS

5.1 Summary of principal findings

5.1.1 Identification of studies

The overall research review question for this review is as follows:

What is the effect of grammar teaching in English on 5 to 16 year olds' accuracy and quality in written composition?

Within this, the research review question identified for the in-depth review is as follows:

What is the effect of teaching *sentence combining* in English on 5 to 16 year olds' accuracy and quality in written composition?

5.1.2 Mapping of all included studies

Twenty-five reviews and 34 primary research studies met the inclusion criteria developed for the overall research review. These reviews and studies were keyworded and formed the basis of the systematic map. The map revealed a number of characteristics of research on the teaching of grammar.

Research reviews

- Thirty-six percent (n = 9) of the included reviews were conducted in the five-year period between 1976 and 1980.
- Two of the reviews are systematic and 23 are non-systematic.
- Sixty-three percent (n = 16) of the reviews were conducted in the USA.
- The majority (60%) of reviews are published.
- Almost all the reviews (n = 24) report on the teaching of sentence-level grammar and the majority of this group (n = 20) focused on sentence combining.

Primary research

- Just under half (n = 15) of the primary studies included in the map were conducted in the ten-year period between 1976 and 1985.
- Almost all the studies (n = 32) report researcher-manipulated evaluations.
- Twenty-two of the studies report trials, of which nine report randomised controlled trials.
- Eighty-two percent (n = 28) of the studies were conducted in the USA.
- The majority (59%) of the studies are published.
- Twenty-nine studies report on sentence-level grammar teaching of which 18 focused on sentence combining.

5.1.3 Nature of studies selected for in-depth review

Eighteen studies met the inclusion criteria for the in-depth review. Table 5.1 summarises the study type and the overall weights of evidence (WoE) assigned to each of these studies.

Table 5.1: Overall weights of evidence assigned to studies

Author, year, title	Overall WoE
Randomized controlled trials	
O'Hare (1973)	High
Saddler and Graham (forthcoming)	High to medium
Stoddard and Renzulli (1983)	Medium
Vitale *et al*. (1971)	Medium
McAfee (1981)	Low to medium
Controlled trials	
Combs (1976,1977)	Medium to high
Hunt and O'Donnell (1970)	Medium to high
Combs and Wilhelmsen (1979)	Medium
MacNeill (1982)	Medium
Mellon (1969)	Medium
Miller and Ney (1967,1968)	Medium
Pederson (1978)	Medium
Melvin (1980)	Medium to low
Nutter and Safran (1983)	Medium to low
Pre-and post-test	
Rousseau and Poulson (1985)	Medium to low
Ney (1976)	Low
Roberts and Boggase (1992)	Low
Rousseau (1989)	Low

From the above table, it can be seen that only four studies – those by O'Hare (1973), Saddler and Graham (forthcoming), Combs (1976, 1977), and Hunt and O'Donnell (1970) – were rated as providing medium to high weight of evidence or above. These studies provide the best evidence in answer to the research question. Whether we look at them chronologically or in terms of their study type, they suggest a positive effect of sentence combining on writing accuracy and quality (specifically, syntactic maturity). The first two carry more weight because of the nature of the studies – randomised controlled trials – in relation to the particular research question we set ourselves: what is the evidence for the effectiveness of grammar teaching (sentence combining) on the accuracy and quality of written composition for 5–16 year olds in English?

5.1.4 Synthesis of findings from studies in in-depth review

An overall synthesis of the results from the eighteen studies examined in the in-depth review comes to a clear conclusion: that sentence combining is an effective means of improving the syntactic maturity of students in English between the ages of 5 and 16. All but two of the studies specify the age group they worked with: predominantly, this group ranged from fourth grade (9–10 year olds) to tenth grade (15–16 year olds), with the majority clustering in the upper years of primary/elementary schooling and the lower years of secondary schooling. The differences between the studies are largely inherent in the *degree* of advance that students learning sentence combining enjoy in terms of their syntactic maturity. In the most reliable studies, immediate post-test effects are seen to be positive with some tempering of the effect in delayed post-tests.

In other words, as might be expected, gains made by being taught sentence combining in terms of written composition are greatest immediately after the intervention and tail off somewhat thereafter. Significantly, in the one study that undertakes a delayed post-test, syntactic maturity gains are maintained, albeit less dramatically than immediately after the event.

The synthesis of results is a narrative one, based on the fact that seventeen of the eighteen studies show a positive effect for sentence combining. The one study that runs counter to the general trend is that by MacNeill (1982) which shows no particular effect of the intervention of a 'Sentencecraft' programme, based on O'Hare's work and as applied to argumentative writing. Unfortunately, although the study as a whole was rated medium, the reporting of the study does not provide sufficient detail to mount a counter-argument to the general results of the synthesis. Further research might wish to replicate MacNeill's and others' studies.

5.2 Strengths and limitations of this systematic review

Of all the systematic reviews so far completed by the English Review Group (Andrews *et al*., 2002; Torgerson and Zhu, 2003; Andrews *et al*., 2004; Burn and Leach, 2004; Locke and Andrews, 2004; Low and Beverton, 2004) this particular review's results are the most emphatic. In no other review have the results pointed to such a clear conclusion: that sentence combining has a positive effect on writing quality and accuracy in terms of syntactic maturity. The initial searching, screening and mapping covered a large field. The particular studies that emerged for the in-depth review have been distilled by careful procedures, as in all systematic reviews. We can be reasonably sure that the present review, taken with its complementary in-depth review on the effectiveness of the teaching of formal grammar (syntax) on written composition, is comprehensive; perhaps, taken together, these form the most comprehensive study of its kind to cover twentieth- and twenty-first century studies of the effectiveness of grammar teaching on written composition.

It is also the case that the principal findings, supported by all but one of the rest of the studies, are based on four studies that were rated medium-to-high or above. The validity and reliability of these studies is strong.

As in all research, there are limitations. These are fully acknowledged and listed as follows:

- Interestingly, the 17 studies that show a positive effect were all conducted in the USA; the one that showed no positive effect was conducted in Canada. There is a question as to how generalisable the results will be in relation to countries outside the USA.
- The majority (14) of the studies in the in-depth review were published in the 1970s and 1980s; two were published in the 1960s; and one each in the 1990s and 2000s. As discussed in the implications sections below, the clustering of studies in two decades does not mean to say that the results are not relevant to current practice and policy.
- We have asked a question about effectiveness; such a tight focus can be seen as a limitation, although it is also a strength in terms of what we can report in relation to this question.

- We were unable to investigate the possibility of the presence of publication bias in the review because we only had four high quality trials with effect sizes in our in-depth review and this is too few for a funnel plot. However, it seems probable that our results are affected by publication bias, given that the two 'unpublished' studies have smaller effect sizes than the two published studies, and the fact that we excluded unpublished PhD theses.

5.3 Implications

5.3.1 Policy

There are a number of implications for policy. We will confine our comments to the implications for policy in England, as other countries may have different policy orientations on the teaching of grammar. However, we expect that there might be points of interest for curriculum design and policy implementation in other countries. We should also point out that the policy documents that we draw on were largely published in the late 1990s; they form the basis of current practice in schools in England, but in our view, with respect to the effectiveness of grammar teaching, they are in need of revision.

First, the National Curriculum for England at Key Stage 1 (5–7 year olds) continues to insist that 'pupils should be taught some of the grammatical features of written standard English' (DfEE, 1999, p 21), and furthermore, in composing their own texts, pupils should be 'taught to consider a) how word choice and order are crucial to meaning, b) the nature and use of nouns, verbs and pronouns, c) how ideas may be linked in sentences and how sequences of sentences fit together' (ibid). Bearing in mind the results and implications of the other in-depth review (Andrews *et al.*, 2004) emerging from our systematic review of the teaching of grammar and its effect on written composition, two of these requirements now look in need of revision: the National Curriculum does not specify what grammatical features should be taught and why; and the twentieth century emphasis on the nature and use of nouns, verbs and pronouns looks vestigial. However, the other two requirements continue to remain important to composition in that they emphasise the actual craft of writing rather than the meta-language of completed sentences.

At Key Stage 2 (7–11 year olds), the requirements become more specific: pupils 'should be taught word classes and the grammatical functions of words, including nouns, adjectives, adverbs, pronouns, prepositions, conjunctions, articles' as well as 'the grammar of complex sentences, including clauses, phrases and connectives' (ibid, p 29). At the next stage (11–14 year olds), pupils should be taught 'the principles of sentence grammar' and 'use this knowledge in their writing". As far as sentence construction goes, they should be taught 'word classes or parts of speech and their grammatical functions', 'the structure of phrases and clauses and how they can be combined to make complex sentences [for example, coordination and subordination] ' and 'the use of appropriate grammatical terminology to reflect on the meaning and clarity of individual sentences' (ibid, p 38).

The results of our two in-depth reviews are that traditional grammar teaching, based on word classes and the teaching of syntax (using a meta-language to

describe and classify parts of speech), is largely ineffective and that sentence combining is largely effective. It follows that much of what is prescribed above for the teaching of writing in the National Curriculum for the early 2000s is highly questionable. We are not implying that traditional top-down grammar teaching is of no use; simply that it does not help young people in learning to write well. Neither are we implying that sentence combining, in all its various forms, is a panacea in helping young people to write well; simply that it has been proved to work and should be considered as an important element in a repertoire of activities, especially for 7–14 year olds, where most of the research has been conducted.

A number of policy initiatives were instigated to support the teaching of the National Curriculum. Most notable have been the Literacy, Numeracy and Key Stage 3 [11–14] National Strategies. An evaluation of the first year of the pilot strategy noted that while there had been improvements in word- and sentence-level work, 'improvements were least in sentence construction, punctuation and paragraphing' (Ofsted, 2002, p 13). There could be a number of factors that have contributed to this; but it could also be the case that, despite increased structuring of the literacy curriculum and timetable (for example, via three-part lessons devoted to literacy), no significant effect occurred with regard to sentence construction. The late 1990s had seen two publications on the teaching of grammar in the National Curriculum, mentioned already in the background to the present report: *The grammar papers: perspectives on the teaching of grammar in the national curriculum* (QCA 1998b) and *Not whether but how: teaching grammar at Key Stages 3 and 4* [for 11–16 year olds] (QCA, 1999). The first of these concludes that 'there are no easy answers to the relationship between knowing how language works and being able to apply that knowledge [and that] it is probably time to shift the criterion by which the usefulness of grammar is judged' (p 55):

> It may be more profitable to promote the teaching of grammar on different grounds: as a strand in the teaching and learning of language, which like all other aspects, compositional and structural, does not have a straight transfer into writing.

Thus much is acknowledged, but there is a vestigial trace of formal grammar in such statements – almost a nostalgia for such teaching. The writers of the document seem unable, or unwilling, to sever ties with a practice that has been proved to be ineffective. Instead, they look for some other justification:

> The routine discussion and teaching of language, including syntactic structures and rules, as part of preparation for and feedback from writing is something which seems to have been lost. In the absence of other evidence, it is this, invigorated with more recent knowledge from linguistics, genre and discourse theories, and a basic core of terminology, which offers the most fruitful way forward (1998, p 56).

While not wishing to decry the value of knowledge about language, it appears that such positions are weak in relation to composing in writing. If the 'fruitful way forward' is framed in questions of *not whether but how*, it would appear that the key question of whether and what to teach has been side-stepped. The myriad of practical approaches to the teaching of grammar – many of them inventive and useful if their own right – tends to mask the fact that very few of these activities are likely to be effective in the teaching and learning of writing.

The main implication for policy of the current review, then, is that the National Curriculum and accompanying guidance needs to be revised to take into account the findings of research: that the teaching of formal grammar (and its derivatives) is ineffective; and the teaching of sentence combining is one (of probably a number of) method(s) that is effective.

5.3.2 Practice

We could characterise current practice in England, within the frameworks established and required by the National Curriculum, to be the use of a range of approaches to traditional grammar, language awareness, the development and use of meta-language to describe sentences and sentence construction itself. The implications of the current in-depth review, taken in concert with its associated review of the teaching of syntax and formal grammar (Andrews *et al.*, 2004), are that the fourth of these is useful and is likely to be effective.

These are important implications, because much current practice in primary and secondary schools in England might be said to be interesting and valuable in its own right, but ineffective in terms of helping young people to write. What the review of research literature in the field has shown is that sentence combining *is* effective.

Following the policy papers discussed above (QCA, 1998b, 1999), the National Literacy Strategy in England issued an extensive teaching manual and accompanying CDRom and video, *Grammar for Writing* (DfEE, 2000), which embodied the principles outlined in the policy papers and was aimed at Key Stage 2 (7–11 year olds). In addition to a wide range of pedagogic activities designed to enliven the teaching of grammar, techniques of expansion (p 44), reduction (p 58) and re-ordering (p 90) are used. There is, however, still a prevailing use of classification, boxing and construction, emphasizing the taxonomic and hierarchical nature of grammar (with its attendant terminology) rather than a more 'horizontal' emphasis on sentence combining. One of the implications of the present review is that *Grammar for Writing* might be revised to strengthen the horizontal dimension and extended to Key Stages 1 (5–7) and 3 (11–14) with appropriate materials.

Sentence combining, as set out in the background section of the report, includes a range of activities in writing, ranging from the actual combining of separate sentences in different ways at one end of the spectrum, to embedding within a single sentence at the other end. The key feature of such pedagogy and practice is that it is *practical*: a hands-on craft activity. As a practice, it needs to be set within meaningful writing contexts, rather than presented as a drill-and-practice exercise.

Further implications for practice which follow are that in-service and pre-service training of teachers in the craft of writing sentences needs to be developed; teachers need to see the business of learning to write from the learners' point of view rather than approach it from the assumption that there is a body of knowledge (unspecified 'recent knowledge from linguistics, genre and discourse theories') that needs to be taught to young people. Our understanding of the current position of teachers at primary and secondary level in England is (a) that they are not secure in such knowledge and thus (b)

in the wake of *Grammar for Writing,* they are likely to take a 'smorgasbord' approach to the teaching of grammar. Some of these approaches will be lively and engaging; others will be abstract and baffling to young people. It is unlikely that all of them are effective in helping young people to write.

A very practical implication of the results of the present review, then, is that it would be helpful if the development of teaching materials and approaches included recognition of the effectiveness of sentence combining.

5.3.3 Research

As ever, more research is needed. However, the two in-depth reviews on the teaching of grammar to improve written composition have enabled us to clear the ground after more than a century of research in the field.

First, we feel we can categorically say that further research into the teaching of formal grammar as an aid to writing is not worth pursuing. We now know that such teaching is ineffective in terms of helping student to write more fluently, more accurately and with more quality. The case has been proven many times and, as far as we know, there is no research of quality that proves otherwise. Researchers should draw a line under that particular field of enquiry and move on.

Second, most of the research on sentence combining we have unearthed took place in the 1970s and 1980s in the USA. There is a need for further studies in other countries. There are aspects of composing – like awareness of genre and textual characteristics, composing with the aid of a word-processor, etc. – that impinge upon it and which require research. We would wish to see future research on sentence combining take into account these and other contextual factors.

Third, if we are to pursue questions of effectiveness, we would suggest that some large-scale studies are undertaken using randomised controlled trial methodologies. To our knowledge, no such studies on the topic in question have taken place outside the USA in the last fifty years or so. A number of such studies, replicated in different contexts and using a longitudinal dimension, would help us to specify with more confidence what works and what does not work.

Fourth, very few of the studies we have examined for either of the two in-depth reviews have been of high quality overall. There is a need for future studies to be well-designed, relevant to the question, and with substantial and well-reported data.

Finally, we do not think that questions of effectiveness are the only research questions that are of interest in terms of the relationship between grammar and writing development. Areas for future research include the examination of the nature of young people's writing as it develops through the school years. There is at least a generation of very good research into emergent grammars and we would like to see the continuation of such a tradition.

6. REFERENCES

6.1 Studies included in map and synthesis

Reviews

Abrahamson RF (1977) The effects of formal grammar instruction vs the effects of sentence combining instruction on student writing: a collection of evaluative abstracts of pertinent research documents. Unpublished research report. Houston, TX, USA: University of Houston.

Amiran E, Mann J (1982) Written composition, grades K–12: literature synthesis and report. Portland, OR, USA: Northwest Regional Educational Laboratory. ERIC document number ED213034.

Asher W (1990) Educational psychology, research methodology, and meta-analysis. *Educational Psychologist* **25:** 143–158.

Bamberg B (1981) Making research work for the composition teacher. Paper presented at the Meeting of the Southland Council of Teachers of English. Los Angeles, CA, USA: October 10. ERIC document number ED208408.

Crowhurst M (1980) The effect of syntactic complexity on writing quality: a review of research. Unpublished research report. British Columbia: University of British Columbia. ERIC document number ED202024.

Elley WB (1994) Grammar teaching and language skill. In: Asher R (ed.) *Encyclopedia of Language and Linguistics*. Oxford: Pergamon (pages 1468–1471).

Gann M (1984) Teaching grammar: is structural linguistics really better? *English Quarterly* **17:** 31–53.

Hillocks G Jr (1984) What works in teaching composition: a meta-analysis of experimental treatment studies. *American Journal of Education* **93:** 133–170.

Hillocks G Jr (1986) *Research on Written Composition: New Directions for Teaching*. Urbana, IL, USA: National Council of Teachers of English. ERIC document number ED265552.

Hudson R (2000) Grammar teaching and writing skills: the research evidence. London: University College London. Available from: http://www.phon.ucl.ac.uk/home/dick/writing.htm

Kolln M (1996) Rhetorical grammar: a modification lesson. *English Journal* **85:** 25–31.

Lawlor J (1980) Improving student writing through sentence combining: a literature review. Technical note. Los Alamitos, CA, USA: Southwest Regional Laboratory for Educational Research and Development. ERIC document number ED192356.

Matzen RN Jr, Zhang W-S, Hartwell P (1995) The role of traditional grammar instruction in the teaching of writing: a selected, annotated bibliography. Paper presented at the Annual Meeting of the Conference on College Composition and Communication. Washington, DC, USA: March 23–25. ERIC document number ED396328.

Newkirk T (1978) Grammar instruction and writing: what we don't know. Close-up: grammar and composition. *English Journal* **67:** 46–48.

Ney JW (1980) A short history of sentence combining: its limitations and use. *English Education* **11:** 169–177.

Phillips SE (1996) Sentence combining: a literature review. Unpublished research report. Dallas, TX, USA: University of Texas. ERIC document number ED398589.

Qualifications and Curriculum Authority (QCA) (1998a) *Recent Research on Grammar Teaching*. Hayes, Middlesex: QCA Publications.

Seidenberg PL (1989) Relating text-processing research to reading and writing instruction for learning disabled students. *Learning Disabilities Focus* **5:** 4–12.

Sternglass MS (1979) Creating the memory of unheard sentences. Paper presented at the Annual Meeting of the South Eastern Conference on English in the Two-Year College. Orlando, Florida, USA: February 8–10. ERIC document number ED176258.

Stewart MF (1979) Sentence-combining: its past, present and future. *English Quarterly* **12:** 21–36.

Stotsky SL (1975) Sentence-combining as a curricular activity: its effect on written language development and reading comprehension. *Research in the Teaching of English* **9:** 30–71.

Tomlinson D (1994) Errors in the research into the effectiveness of grammar teaching. *English in Education* **28:** 20–26.

Ulin RO, Schlerman BJ (1978) For the sake of teaching writing, deliver us from traditional grammar. *High School Journal* **62:** 58–68.

Walsh SM (1991) Breakthroughs in composition instruction methods without evidence of tangible improvements in students' composition: when will change come? Paper presented at the Annual Spring Conference of the National Conference of Teachers of English. Indianapolis, IN, USA: March 14–16. ERIC document number ED336744.

White RS, Karl H (1980) Reading, writing and sentence combining: the track record. *Reading Improvement* **17:** 226–232.

Wyse D (2001) Grammar. For writing? A critical review of empirical evidence. *British Journal of Educational Studies* **49:** 411–427.

Primary research

Aulls MW (2003) The influence of a reading and writing curriculum on transfer learning across subjects and grades. *Reading Psychology* **24:** 177–215.

Bateman DR, Zidonis FJ (1966) *The Effect of A Study of Transformational Grammar on the Writing of Ninth and Tenth Graders*. Urbana, IL, USA: National Council of Teachers of English.

Calkins LM (1979) When children want to punctuate: basic skills belong in context. Paper presented at the Annual Meeting of the National Conference on Language Arts in the Elementary School. Hartford, Connecticut, USA: March 23–25. ERIC document number ED170766.

Calkins LM (1980) When children want to punctuate: basic skills belong in context. *Language Arts* **57:** 567–573.

Combs WE (1976) Further effects of sentence-combining practice on writing ability. *Research in the Teaching of English* **10:** 137–149.

Combs WE (1977) Sentence-combining practice: do gains in judgments of writing 'quality' persist? *Journal of Educational Research* **70:** 318–321.

Combs WE, Wilhelmsen K (1979) In-class 'action' research benefits research, teacher and students. *High School Journal* **62:** 267–271.

Elley WB, Barham IH, Lamb H, Wyllie M (1975) The role of grammar in a secondary school curriculum. *New Zealand Council for Educational Studies* **10:** 26–41.

Elley WB, Barham IH, Lamb H, Wyllie M (1979) *The Role of Grammar in a Secondary School Curriculum. Educational Research Series No 60*. Wellington: New Zealand Council for Educational Research. ERIC document number ED185588.

Fogel H, Ehri LC (2000) Teaching elementary students who speak black English vernacular to write in standard English: effects of dialect transformation practice. *Contemporary Educational Psychology* **25:** 212–235.

Gordon CJ (1990) Contexts for expository text structure use. *Reading Research and Instruction* **29:** 55–72.

Green S, Sutton P (2003) What do children think as they plan their writing? *Reading: Literacy and Language* **37:** 32–38.

Hilfman T (1970) Can second grade children write more complex sentences? *Elementary English* **47:** 209–214.

Holdich CE, Chung PWH, Holdich RG (2004) Improving children's written grammar and style: revising and editing with HARRY. *Computers and Education* **42:** 1–23

Hunt KW, O'Donnell R (1970) *An Elementary School Curriculum to Develop Better Writing Skills*. Washington DC: Office of Education, Bureau of Research.

MacNeill TB (1982) The effect of sentence-combining practice on the development of reading comprehension and the written syntactic skills of ninth grade students. Unpublished research report. Alberta: University of Alberta. ERIC document number ED217415.

McAfee D (1981) Effect of Sentence Combining on Fifth Grade Reading and Writing Achievement. Paper presented at the Annual Meeting of the National Reading Conference, Dallas, TX, USA: December 2–5. ERIC document number ED217388.

McNeill JH (1994) Instruction for deaf students in syntactic cohesion. *Acehi Journal/Revue Aceda* **20:** 88–95.

Mellon J (1969) *Transformational Sentence Combining. Research Report No. 10*. Urbana, IL, USA: National Council of Teachers of English.

Melvin MP (1980) The effects of sentence combining instruction on syntactic maturity, reading achievement, and language arts skills achievement. Unpublished research report. Oxford, OH, USA: Miami University. ERIC document ED191007.

Miller BD, Ney JW (1967) Oral drills and writing improvement in the fourth grade. *Journal of Experimental Education* **36:** 93–99.

Miller BD, Ney JW (1968) The effect of systematic oral exercises on the writing of fourth-grade students. *Research in the Teaching of English* **2:** 44–61.

Ney JW (1976) Sentence combining and reading. Unpublished research report. Tempe, AZ, USA: Arizona State University. ERIC document number ED161080.

Nutter N, Safran SP (1983) Sentence combining and the learning disabled student. Unpublished research report. Athens, OH, USA: Ohio University. ERIC document number ED252994.

O'Hare F (1973) *Sentence Combining: Improving Student Writing without Formal Grammar Instruction. Research Report No 15.* Urbana, IL, USA: National Council of Teachers of English.

Pedersen EL (1978) Sentence-combining practice: training that improves student writing. Unpublished research report. Utah: Weber State College. ERIC document number ED169567.

Roberts CM, Boggase BA (1992) Non-intrusive grammar in writing. Paper presented to the Annual Conference on Computers and Writing. Indianapolis, USA: May 1–3. ERIC document number ED348684.

Robinson CF (1978) An investigation of new rhetoric lessons for improved written composition on the secondary school level. Unpublished research report. Hartford, CT, USA: Hartford Public School System. ERIC document number ED188196.

Rousseau MK (1989) Increasing the use of compound predicates in the written compositions of students with mild learning handicaps. Paper presented at the annual convention of the association for behaviour analysis, Nashville, TN, USA: May. ERIC document number ED342154.

Rousseau MK, Poulson CL (1985) Using sentence-combining to teach the use of adjectives in writing to severely behaviorally disordered students. Unpublished research report. New York: City University of New York. ERIC document number ED342153.

Saddler B, Graham S (forthcoming) The effects of peer-assisted sentence combining instruction on the writing performance of more and less skilled young writers. *Journal of Educational Psychology*.

Satterfield J, Powers A (1996) Write on! Journals open to success. *Perspectives in Education and Deafness* **15:** 2–5.

Stock R (1980) The effect of teaching sentence patterns on the written sentence structures of grade two children. Paper presented at the Annual Meeting of the Plains Regional Conference of the International Reading Association, Bismark ND, USA: September 25–27. ERIC document number ED208414.

Stoddard EP, Renzulli JS (1983) Improving the writing skills of talent pool students. *Gifted Child Quarterly* **27:** 21–27.

Stone AK, Serwatka TS (1982) Reducing syntactic errors in written responses of a retarded adolescent through oral patterning. *Education and Training in Mental Retardation and Developmental Disabilities* **17:** 71–74.

Thompson CL, Middleton M (1973) Transformational grammar and inductive teaching as determinants of structurally complex writing. *California Journal of Educational Research* **24:** 28–41.

Vitale MR, King FJ, Shontz DW, Huntley GM (1971) Effect of sentence-combining exercises upon several restricted written composition tasks. *Journal of Educational Psychology* **62:** 521–525.

Welch M (1992) The PLEASE strategy: a metacognitive learning strategy for improving the paragraph writing of students with mild learning disabilities. *Learning Disability Quarterly* **15:** 119–128.

6.2 Other references used in the text of the report

Andrews R, Burn A, Leach J, Locke T, Low G, Torgerson C (2002) A systematic review of the impact of networked ICT on 5-16 year olds' literacy in English. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education. Available from: http://eppi.ioe.ac.uk/EPPIWebContent/reel/review_groups/english/eng_rv1/eng_rv1.pdf

Andrews R, Torgerson C, Beverton S, Locke T, Low G, Robinson A, Zhu D (2004) The effect of grammar teaching (syntax) in English on 5 to 16 year olds'

accuracy and quality in written composition. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education. Available from: http://eppi.ioe.ac.uk/EPPIWebContent/reel/review_groups/english/eng_rv6/eng_rv6.pdf

Asker W (1923) Does knowledge of formal grammar function? *School and Society* **27th January**: 109–111.

Benfer M (1935) Sentence sense in relation to subject and predicate. Unpublished Master's thesis. Iowa City, Iowa: University of Iowa.

Boraas J (1917) Formal grammar and the practical mastery of English. Unpublished doctoral thesis. Minneapolis, Minnesota: University of Minnesota.

Braddock R, Lloyd-Jones R and Schoer L (1963) *Research on Written Composition*. Illinois, USA: National Council of Teachers of English.

Burn A, Leach J (2004) A systematic review of the impact of ICT on the learning of literacies associated with moving image texts in English, 5-16. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education. Available from: http://eppi.ioe.ac.uk/EPPIWebContent/reel/review_groups/english/eng_rv5/eng_rv5.pdf

Catherwood C (1932) A study of the relationship between a knowledge of rules and ability to correct grammatical errors and between identification of sentences and knowledge of subject and predicate. Master's thesis. Minneapolis, Minnesota: University of Minnesota.

Chomsky N (1957) *Syntactic Structures.* The Hague: Mouton.

Damasio A (2000) *The Feeling of What Happens: Body, Emotion and the Making of Consciousness.* London: Vintage.

Department of Education and Science (DES) (1975) *A Language for Life* (The Bullock Report). London: HMSO.

Department of Education and Science (DES) (1988) *Report of the Committee of Inquiry into the Teaching of the English Language* (The Kingman Report). London: HMSO.

Department for Education and Employment (DfEE) (1999) *The National Curriculum for England: English.* London: DfEE/Qualifications and Curriculum Authority.

Department for Education and Employment (DfEE) (2000) *The National Literacy Strategy: Grammar for Writing*. London: Department for Education and Employment (book and video).

DiStefano P and Killion J (1984) Assessing writing skills through a process approach. *English Education* **16:** 203–207

Dixon RMW (1965) *What is Language? A New Approach to Linguistic Description.* London: Longmans, Green and Co Ltd.

Elley WB, Barham IH, Lamb H, Wyllie M (1979) *The Role of Grammar in a Secondary School Curriculum. Educational Research Series No 60*. Wellington: New Zealand Council for Educational Research. ERIC document number ED185588.

EPPI-Centre (2002a) *Core Keywording Strategy: Data Collection for a Register of Educational Research. Version 0.9.7.* London: EPPI-Centre, Social Science Research Unit.

EPPI-Centre (2002b) *Guidelines for Extracting Data and Quality Assessing Primary Studies in Educational Research. Version 0.9.7.* London: EPPI-Centre, Social Science Research Unit.

EPPI-Centre (2002c) EPPI-Reviewer. Version 2.5.2. London: EPPI-Centre, Social Science Research Unit.

Fairclough N (1992) *Discourse and Social Change.* Cambridge: Polity Press.

Halliday MAK (1988) On the language of physical science. In: Ghadessy M (ed.) *Registers of Written English: Situational Factors and Linguistic Features*. London: Pinter (pages 162–178).

Halliday MAK (1989) Some grammatical problems in scientific English. *Australian Review of Applied Linguistics: Genre and Systemic Functional Studies* **6**: 13–37. Reprinted in: Halliday MAK, Martin JR (eds) *Writing Science: Literacy and Discursive Power*, London: Falmer Press (pages 69–85).

Halliday MAK, Hasan R (1976) *Cohesion in English*. London: Longman.

Halliday MAK, Hasan R (1985) *Language, Context and Text: Aspects of Language In a Social Semiotic Perspective.* Oxford: Oxford University Press.

Harris RJ (1962) An experimental enquiry into the functions and value of formal grammar in the teaching of English, with special reference to the teaching of correct written English to children aged twelve to fourteen. Unpublished PhD thesis. London: University of London.

Hillocks G, Smith MW (1991) Grammar and Usage. In: Flood J, Jensen JM, Lapp D, Squire JR (eds) *Handbook of Research on Teaching the English Language Arts*. New York: Macmillan (pages 591–603).

Hodge R, Kress G (1993) *Language as Ideology*. 2nd edition. London: Routledge.

Hudson R (1992) *Teaching Grammar: A Guide for the National Curriculum.* Oxford: Basil Blackwell.

Kress G (1994) *Learning to Write.* 2nd edition. London: Routledge.

Lakoff G, Johnson M (1980) *Metaphors We Live By*. Chicago: Chicago University Press.

Locke T, Andrews R (2004) A systematic review of the impact of ICT on literature-related literacies in English, 5–16. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education. Available from: http://eppi.ioe.ac.uk/EPPIWebContent/reel/review_groups/english/eng_rv3/eng_rv3.pdf

Low G, Beverton S (2004) A systematic review of the impact of ICT on literacy learning in English of learners between 5 and 16, for whom English is a second or additional language. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education. Available from: http://eppi.ioe.ac.uk/EPPIWebContent/reel/review_groups/english/eng_rv4/eng_rv4.pdf

Macaulay WJ (1947) The difficulty of grammar. *British Journal of Educational Psychology* **17:** 153–162.

Ministry of Education (1996) *Exploring Language: A Handbook for Teachers.* Wellington, NZ: Learning Media.

Ofsted (2002) *The Key Stage 3 Strategy: Evaluation of the First Year of the Pilot*. London: Office for Standards in Education.

O'Hare F (1973) *Sentence Combining: Improving Student Writing without Formal Grammar Instruction. Research Report No 15.* Urbana, IL, USA: National Council of Teachers of English.

Perera K (1984) *Children's Writing and Reading: Analysing Classroom Language.* Oxford: Basil Blackwell.

Pinker S (1995) *The Language Instinct: The New Science of Language and Mind.* London: Penguin.

Qualifications and Curriculum Authority (1998b) *The Grammar Papers: Perspectives on the Teaching of Grammar in the National Curriculum*. London: QCA

Qualifications and Curriculum Authority (1999) *Not Whether But How: Teaching Grammar in English at Key Stages 3 and 4*. London: QCA

Quirk R, Greenbaum S, Leech G, Svartvik J (1985) *A Comprehensive Grammar of the English Language.* London: Longman.

Rice JM (1903) Educational research: the results of a test in language and English. *Forum* **XXXV:** 209–293, 440–457.

Robinson N (1960) The relationship between knowledge of English grammar and ability in English composition. *British Journal of Educational Psychology* **30:** 184–186.

Saddler B, Graham S (2004) The effects of sentence combining practice on writing skills. Paper given at the AERA annual conference, San Diego, USA, April 12–16.

Segal D, Barr NR (1926) Relation and achievement in formal grammar to achievement in applied grammar. *Educational Research* **14:** 401–402.

Symonds PM (1931) Practice vs. grammar in the learning of correct usage. *Journal of Educational Psychology* **22:** 81–96.

Torgerson C, Zhu D (2003) A systematic review and meta-analysis of the effectiveness of ICT on literacy learning in English, 5–16. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education. Available from: http://eppi.ioe.ac.uk/EPPIWebContent/reel/review_groups/english/eng_rv2/eng_rv2.pdf

Wilkinson A (1971) *The Foundations of Language: Talking and Reading in Young people.* London: Oxford University Press.

Wyse D (2001) Grammar. For writing? A critical review of empirical evidence. *British Journal of Educational Studies* **49:** 411–427.

APPENDIX 1.1: ADVISORY GROUP STRUCTURE

The EPPI English Advisory Group

James Durran, Parkside Community College, Cambridge
Polly Griffith, Chair of Governors, Millthorpe School, York
Nick McGuinn, Department of Educational Studies, University of York
Gloria Reid, Kingston-upon-Hull Learning Services
Peter Taylor, All Saints and Oaklands Schools
Ian Watt, Department of Health Sciences, University of York

Literature searching, information management and administrative support

Alison Robinson, University of York

The role of the Advisory Group/user involvement

Meetings are planned twice-yearly to discuss and guide the work of the review group.

In addition to members of the Advisory Group, other users of the research were consulted on the draft protocol, at mapping stage and when a draft of the final report is ready. The Department of Educational Studies is developing its links with schools interested in research in 2003/04 (see Department Plan, available from Alison Robinson). Such links will enable more teachers than those on the advisory group to comment on, contribute to and disseminate the work of the English Group. In addition, following a meeting with the Teacher Training Agency and PGCE students in June 2003, PGCE tutors and students will be involved in a pilot project to write summaries of the present research review (and previous reviews) and to prepare sample lessons arising from the research findings. The dissemination strategy of the English Review Group was discussed at the steering group meeting in September 2003.

In addition, a pupil from Millthorpe School, York, will work on a pupil summary of the final review.

APPENDIX 2.1: INCLUSION AND EXCLUSION CRITERIA

For a paper to be included in the systematic map, it will have to be a study looking at the effect of grammar teaching in English on 5 to 16 year olds' accuracy and quality in written composition. As the focus of the study is on the *effects* of grammar teaching, papers using methods to identify any such effects are required. This implies the following study types, classified according to the EPPI-Centre taxonomy of study type contained in its core keywording strategy (EPPI-Centre 2002a):

B: Exploration of relationships
C: Evaluation (naturally-occurring or researcher-manipulated)
E: Review (systematic or other review) containing at least one study exploring relationships or one evaluation

Inclusion criteria

- Must be a study of the effects of grammar teaching on writing
- Must focus exclusively on children and young people aged 5 to 16
- Must be in a mainstream school setting
- Must be one of the following study types: B (exploration of relationships); C (evaluation); E (review)
- Must be published or unpublished (but in the public domain) between 1900 and the present
- Must be of teaching of English grammar in an English-speaking country
- Must be of teaching of English as first language, not foreign or second or additional language

Exclusion criteria

EXCLUSION ON SCOPE

One:	Not grammar teaching
Two:	Not children or young people aged between 5 and 16
Three:	Not effects of grammar teaching on writing
Three (a):	Not teaching of English grammar (syntax) in an English-speaking country

EXCLUSION ON STUDY TYPE

Four:
(a) A (description)
(b) D (methodology)
(c) Editorial, commentary, book review
(d) Policy document
(e) Resource, textbook
(f) Bibliography
(g) Dissertation abstract
(h) Theoretical paper
(i) Position paper

EXCLUSION ON SETTING IN WHICH STUDY WAS CARRIED OUT

Five:	English as a foreign, second or additional language (L2, EFL, ESL, EAL)
Six:	Not mainstream school setting

Seven:	The effects of grammar teaching on the writing of pupils in a foreign language (e.g. Hebrew, Dutch)

APPENDIX 2.2: ELECTRONIC SEARCH STRATEGY

Review question

What is the effect of grammar teaching on the accuracy and quality of 5 to 16 year olds' written composition?

Databases searched

ERIC

1. exp *grammar/ or exp *syntax/
2. exp *sentence structure/
3. *writing (composition)/
4. *metalinguistics/
5. *cohesion (written composition)/ or *generative grammar/ or *sentence combining/ or *sentence diagraming/ or *structural grammar/ or *text structure/ or *traditional grammar/
6. *case (grammar)/ or *grammatical acceptability/
7. *transformational generative grammar/
8. *coherence/ or *paragraph composition/
9. "KAL".mp
10. 1 or 2 or 3 or 4 or 5 or 6 or 6 or 8 or 9
11. limit 10 to English language
12. limit 11 to (elementary secondary education or elementary education or primary education or intermediate grades or secondary education or middle schools or junior high schools or high schools or high school equivalency programs)
13. limit 12 to (books or conference proceedings or dissertations or "evaluative or feasibility reports" or general reports or information analyses or journal articles or "research or technical reports" or "speeches or conference papers")

PsycINFO

1. ("grammar-" in DE) or ("transformational-generative grammar" in DE)
2. "syntax-" in DE
3. "sentence-structure" in DE
4. "text-structure" in DE
5. (writ*) and (composition*)
6. "metalinguistics-" in DE
7. 1 or 2 or 3 or 4 or 5 or 6
8. limit 7 to ((AG:PY = Adolescence) or (AG:PY = childhood) or (AG:PY = school age)) and (LA:PY = English) and ((PT:PY = case-study) or (PT:PY = conference-proceedings-symposia) or (PT:PY = empirical-study) or (PT:PY = followup-study) or (PT:PY = journal-abstract) or (PT:PY = journal information) or (PT:PY = journal-review-book) or (PT:PY = literature-review-research-review) or (PT:PY = meta-analysis) or (PT:PY = prospective-study) or (PT:PY = retrospective-study) or (PT:PY = treatment-outcome-study))

SSCI

(((gramma* or synta* or sentence structure or metlinguistic* or knowledge about language or KAL))) and ((writ* or composition*)) and (child* or adolescen* or school* or education*)
Doc type = all document types
Language = English

APPENDIX 2.3: EPPI-Centre Keyword sheet

1. Identification of report
Citation
Contact
Handsearch
Unknown
Electronic database
(Please specify.) ……

2. Status
Published
In press
Unpublished

3. Linked reports
Is this report linked to one or more other reports in such a way that they also report the same study?

Not linked
Linked (Please provide bibliographical details and/or unique identifier.)
……
……
……
……

4. Language (Please specify.)

……

5. In which country/countries was the study carried out?
(Please specify.)

……

6. What is/are the topic focus/foci of the study?
Assessment
Classroom management
Curriculum
Equal opportunities
Methodology
Organisation and management
Policy
Teacher careers
Teaching and learning
Other (Please specify.)……

7 Curriculum
Art
Business studies
Citizenship
Cross-curricular
Design and technology
Environment
General
Geography
Hidden
History
ICT
Literacy – first language
Literacy further languages
Literature
Maths
Music
PSE
Phys. Ed.
Religious Ed.
Science
Vocational
Other (Please specify.) ……

8. Programme name (Please specify.)

……

9. What is/are the population focus/foci of the study?
Learners*
Senior management
Teaching staff
Non-teaching staff
Other education practitioners
Government
Local education authority officers
Parents
Governors
Other (Please specify.) ……

10. Age of learners (years)
0–4
5–10
11–16
17–20
21 and over

11. Sex of learners
Female only
Male only
Mixed sex

12. What is/are the educational setting(s) of the study?
Community centre
Correctional institution
Government department
Higher education institution
Home
Independent school
Local education authority
Nursery school
Post-compulsory education institution
Primary school
Pupil referral unit
Residential school
Secondary school
Special needs school
Workplace
Other educational setting ……..

13. Which type(s) of study does this report describe?

Description
Exploration of relationships
Evaluation
    Naturally-occurring
    Researcher-manipulated*
Development of methodology
Review
    Systematic review
    Other review

*see 14.

14. To assist with the development of a trials register please state if a researcher-manipulated evaluation is one of the following:

a. Controlled trial (non-randomised)
b. Randomised controlled trial (RCT)

Please state here if keywords have not been applied from any particular category (1–10) and the reason why (e.g. no information provided in the text).

……
……
……
……
……
……

APPENDIX 2.4: Review-specific keywords

1. On what 'type' of grammar teaching does the study focus? a. 'text' level grammar teaching b. 'sentence' level grammar teaching	2. If 'text' level, is the focus on: a. text structure? b. cohesion? c. coherence? d. paragraph composition? e. not applicable	3. If 'sentence' level, is the focus on: a. syntax? b. sentence-diagramming? c. sentence combining? d. punctuation? e. not applicable?	4. What 'type' of intervention does the study involve? a. contextualised grammar teaching b. decontextualised grammar teaching
5. On what kind of grammar does the study focus? a. language-awareness b. meta-language c. traditional grammar d. transformative/generative grammar e. 'functional' grammar f. 'pedagogic' grammar	6. What 'type' of written outcomes are reported? a. accuracy of writing (please specify) b. quality of writing (please specify)	7. What measurements are reported? a. test results (please specify) b. examination results (please specify) c. written composition (please specify) d. other (please specify)	8. What are the specific characteristics of the learners? a. learning difficulties b. specific learning difficulties (dyslexia) c. other (please specify) d. not applicable

For definitions, see Glossary.

APPENDIX 4.1: Summary tables for studies included in the in-depth review

1. Combs WE (1976) Further effects of sentence-combining practice on writing ability. 2. Combs WE (1977) Sentence-combining practice: Do gains in judgments of writing "quality" persist?	
Country of study	USA
Age of learners	12–13: Seventh grade
Type of study	Researcher-manipulated evaluation: controlled trial (cluster)
Aims of study	The 1976 study aimed to test the following hypotheses: • Syntactic maturity gains achieved by the Mellon and O'Hare procedures [using sentence-combining techniques] are replicable with a seventh-grade population. • Syntactic maturity gains are retained as measured by a delayed post-test of students' free writing. • The overall 'quality' of writing of students receiving SC practice will be judged superior to that of students not receiving SC practice as measured by an expanded matched-pairs design (p 138). The 1977 study added the following hypothesis: • Differentiated levels in quality of writing are retained as measured by a delayed post-test (p 320).
Summary of study design, including details of sample	• 'The design of the study included two intact experimental classrooms and two intact control classrooms selected from a suburban Minneapolis junior high school and followed the pre-test control group design....excepting the random selection of the student population and the inclusion of a delayed post-test' (1976, p 138). • Sample number = 100
Data-collection instruments, including details of checks on reliability and validity	• The nature of the data-collection process is set out on p 139 (1976) beginning 'All students completed the same number and kind of writing exercises regardless of group assignment...' • Further data are provided by teacher-raters who assessed quality writing using matched pairs. • Reliability is addressed, by emphasising that all students 'completed the same number and kind of writing exercises regardless of group assignment'...all students 'studied mythology to the extent required by individual contracts' and so on. • In the assessment of writing quality, seven teacher-raters were used who did not have intimate knowledge of the O'Hare study. All seven rated the matched pairs of assignments. Interventions were based on O'Hare's examples, so tried and tested. • Validity is addressed in that the data collected was clearly related to what was being measured, i.e. writing ability. The author also notes: 'Narrative and descriptive mode were selected since they seem the most typical of junior high writing' (1976, p 140). • The teacher-raters assessing the matched pairs of compositions 'were encouraged to make a single intuitive judgment of the relative quality of each pair of compositions' (1976, p 141).
Methods used to analyse data, including details of checks on reliability and	• T-test comparisons allowing for a comparison of mean change scores for syntactic maturity. Statistical analysis was also used to measure changes in quality of writing. 'Each composition that was checked by a rater received one point, the other zero. Thus, each student composition received a score from zero to seven, depending upon

validity	the number of raters that judged it superior. The mean scores of the students in each group were then computed and compared at both pre- and post-test' (1976, p 145). • There is no indication that the researcher had someone else duplicate the analysis. However, he does cite other researchers who have developed the methods he has adopted; that is, he appeals to the authority of previous usage. • The author justifies the use of calculation of W/TU and W/C 'since there is ample indication that these two indices are the most discriminating of those examined to date' (1976, p 140) citing previous research (Hunt, 1965, p 1970) to back up this claim.
Summary of results	All four research hypotheses were confirmed, '...but in ways interesting enough to invite considerable comment'. In particular, comparisons with previous studies indicate variation in the magnitudes of growth. • Hypothesis 1 was clearly confirmed, with considerable variation in the results of the three studies (Mellon, O'Hare and the present one) in terms of the extent of the growth. Combs finds a grade leap of + 2, as opposed to Mellon (+ 1) and O'Hare (+ 5) years/grades. However, this may be because the experimental period was shorter for Combs than for O'Hare. • Hypothesis 2 was confirmed: specifically, significant difference between the control and experimental groups was sustained, although 'about half the gain (in W/TU) is eroded eight weeks following treatment when no further SC instruction is permitted'. • Hypothesis 3 was confirmed: 'The teacher-raters judged the compositions of the experimental group significantly better as a result of the SC treatment' (p 145). The 1977 study added the following hypothesis: • Differentiated levels in quality of writing are retained as measured by a delayed post-test (p. 320). This hypothesis was confirmed by this study.
Conclusions	Conclusions are no different from results in this study. Put simply, the use of SC exercises was found to be effective in enhancing seventh graders' syntactic maturity by two grades; such advances were sustained and writing quality improved also. • 'The results allow the conclusion that the experimental group wrote compositions that were (1) syntactically more mature than those of the control group and (2) syntactically more mature than those they had written at pre-test' (1976, p 146). • 'It can be concluded the SC practice has a positive effect on the judged "quality" of writing of seventh graders in conjunction with gains in syntactic maturity levels of sentences appearing in those students' compositions. Delayed testing shows a decrease in syntactic maturity levels which do continue to distinguish significantly between control and experimental groups' (1976, p 147). • 'SC practice seemed to affect more than syntactic gains, indeed, gains that were incorporated in what teacher-raters consider improved quality of writing...The present findings provide obvious encouragement for the inclusion of SC activity in the language arts curriculum' (1977, p 321).

Weight of evidence A (trustworthiness in relation to study questions)	High
Weight of evidence B (appropriateness of research design and analysis)	Medium
Weight of evidence C (relevance of focus of study to review)	High to medium
Weight of evidence D (overall weight of evidence)	Medium to high

Combs WE, Wilhelmsen K (1979) In-class 'action' research benefits research, teacher and students.	
Country of study	USA (Madison County, Georgia)
Age of learners	13–14: Eighth grade
Type of study	Researcher-manipulated evaluation: controlled trial
Aims of study	The study aims at making 'research results [in sentence combining as helping writing] more accessible to classrooms' by trialling SC as a classroom intervention in an actual classroom as part of their Language Arts programme. The implication is that the research aim is to look at an application of sentence-combining instruction and its effect on greater writing fluency and increased syntactic maturity; a secondary (curriculum) aim is to 'provide a well-rounded language arts program for the students' (p 268).
Summary of study design, including details of sample	• This is a pre-test, post-test study with a number of delayed post-tests (at seven, fourteen, twenty weeks after the experimental period, and then at one year after it). The sample is not randomised. It is a clustered trial. • SC pedagogy (including revision/review) was applied to the experimental group, with both groups being assessed for syntactic maturity using pre-test and post-test measures. • Both control and experimental groups appear to be single classes, which suggest that these are relatively small middle schools. Size information is not provided.
Data-collection instruments, including details of checks on reliability and validity	• The precise nature of the writing tasks (beyond mode) and who collected these pieces is not spelled out. • The issue of reliability is not discussed. Reliability *is* an issue, in that one would want to assume, for instance, that the conditions under which writing assignments were set were consistent between the two groups. • Validity is discussed only in that the exercises were designed to suit the context and be fun within 'a well-rounded language arts program'. The emphasis on collecting writing in two modes suggests a desire to have comparability between both classes in terms of *types* of writing being produced and collected as data.
Methods used to analyse data, including details of checks on reliability and validity	• Presumably, the calculation of T-unit values. • The mean T-unit lengths were calculated for control and experimental groups. Statistical significance was also used. • The issue of reliability is not directly addressed. We could say that there was longitudinal reliability, in that tests were conducted on several occasions after the end of the experimental period; but that's hardly analytical reliability. No doubt both teacher and researcher were involved in the analysis, but again, not very reliably, one imagines. However, sub-samples were selected randomly at the 7-week and 14-week points to test the reliability of the analysis, so there is some degree of reliability in that. • The validity of W/T-unit analysis as a measure of syntactic maturity is not questioned.
Summary of results	The results reveal that 'our experimental students showed gains in syntactic maturity between the pre-test and post-test writings, particularly in their argumentative papers' underscoring the belief that the argumentative mode lends itself to more complex sentences. But advances in narrative writing were also demonstrated. A year after the end of the experimental period, there was still a significant difference between the experimental group and the control group. 'The experimental papers averaged 12.97 w/T-unit and the control papers 10.71 (narrative mode)' (p 269). However, the major syntactic gains appeared during the treatment period.

	Specifically: • 'Our experimental group students showed gains in syntactic maturity between the pretest and postest writings, particularly in their argumentative papers' (p 269). • Sub-groups of the experimental group who had subsequent review/revision 'boosting' after the experimental period tended to sustain gains made in syntactic maturity. • One year after the intervention, levels of syntactic maturity between control and experimental groups differed significantly in favour of the experimental group.
Conclusions	• The authors conclude that action research is a useful methodology which has indicated firmly that SC techniques that have been shown to work in carefully controlled situations 'do indeed translate into the English classroom' (p 270). However, it is worth pursuing 'action research' as the authors call it – perhaps more accurately described as 'applied research in classrooms'. • There are other conclusions of note and of interest. For instance, the authors conclude that the more impressive gains in syntactical maturity (for the experimental group) in the argumentative mode 'underscores the belief that the argumentative mode lends itself to more complex syntax' (p 269). • Another conclusion interprets the study findings in the light of the fact that the control group was taught 'formal grammatical knowledge'. They conclude: '...we...submit that if one wants students to do well on grammar tests, teach them grammar. But do not expect it to have measurable influences on their writing maturity. However, if one wants the grammatical structures to show in students' free writing, then present them SC strategies in systematic, extended, and creative lessons' (p 270). • One should note that this study had, perhaps inadvertently, provided a finding of relevance to the topic of this project's previous review.
Weight of evidence A (trustworthiness in relation to study questions)	Medium
Weight of evidence B (appropriateness of research design and analysis)	Medium As a controlled trial pre- and post-test study with a number of delayed post-tests, this design is fairly appropriate; but the sample is not randomly allocated. As a clustered trial, it provides medium weight of evidence.
Weight of evidence C (relevance of focus of study to review)	High *Conceptual focus:* The conceptual focus (on sentence combining, syntactical maturity, writing in context, generative grammar and language in use) is highly relevant. *Context:* The focus on a particular English classroom using an action research model drawing on previous research makes this context highly relevant (as evidenced by the response of local English teachers). *Sample:* The sample (a "normal" class of eighth grade) is relevant by virtue of its ordinariness. *Measures:* The measures of syntactic maturity are clearly appropriate and relevant to the research question.

Weight of evidence D (overall weight of evidence)	Medium Overall, the study has provided medium weight of evidence, largely because there is insufficient contextual description and non-randomisation of the sample. More rigour, either in the execution or the reporting, would have given this study a higher weight of evidence.

Hunt KW, O'Donnell R (1970) An elementary school curriculum to develop better writing skills	
Country of study	USA
Age of learners	9–10: Fourth grade
Type of study	Researcher-manipulated evaluation: controlled trial (cluster)
Aims of study	To examine the impact of sentence combining on the writing of fourth grade students, specifically: 'to try out a sentence-combining curriculum in the fourth grade for black and white students and then to test its effect on ... syntactic maturity' (p 6).
Summary of study design, including details of sample	This is a cluster trial design with (a) a pre-test, mid-term test and post-test, but no delayed post-test, and (b) two experimental groups and two control groups. • Pre-test: free writing, rewriting test, Nelson reading test • Intervention 1: mid-term test – free writing and rewriting tests from pre-test • Intervention 2: post-test – free writing test, different rewriting test, Stedman Reading Structure Test • N = 335, comprising 194 black children and 141 white children. Experimental group N = 180, control group N = 155.
Data-collection instruments, including details of checks on reliability and validity	'The pre-tests were in the general area of writing and a special kind of rewriting and reading...the post-test consisted of not just one but of three pieces of writing' (p 9). • Free writing test: a 'short silent cartoon', then students were 'asked to tell what the movie was about'. Pre-test involved one free writing exercise. Post-test involved three such exercises. • Rewriting test: students were given text composed of 28 very short simple sentences. They had to rewrite 'in a better way' (p 10). Better was defined by the researchers as creating longer sentences by combining the shorter ones. It is implied that the students were not told this. • Nelson reading test: Revised ed., grades 3–9, 175 items comprising 100 vocabulary items (not scored) + 75 comprehension items (scored). • Stedman Reading Structure Test: students were given gapped passages and a list of structure words to add in. The number of gaps is not reported, nor whether the list had more words than the number of gaps. • Reliability addressed via a mid-test: 'a correlation of those control students' scores on the pre- and mid-term tests was made as an indication of the reliability of the instruments' (p 12). The post-test rewriting exercise had been used previously. • Validity was addressed by extending the post-test (but not the pre-test) to three writing texts in order to increase the number of words. Only the reading comprehension scores on the Nelson Test were used as an index of reading level. The authors are at least honest about stating that they did not know what precisely the modified cloze test was testing.
Methods used to analyse data, including details of checks on reliability and validity	• Words per T-unit, clauses per T-unit and words per clause, plus sentence embeddings (via tree diagrams) • Mean scores, correlations, standard deviations and analysis of covariance • Reliability was addressed by increasing the number of free writing tests in the post-test. The results of the three tests are not reported separately or compared. Reliability was further addressed by using ANCOVA.

	• Validity was addressed by checking for IQ or reading skill biases in the sample. The experimental group had a higher mean IQ (at 0.05 level), but there was no reading skill bias. The researchers were concerned to cross-check the type of embedding in the rewriting tests against previous findings, as an indication of robustness. This applied particularly to the ability to reduce a clause to a single word (e.g. adjective), which they felt was a good indicator of later syntactic development (at twelfth grade) (p 20).
Summary of results	• 'At the end of the year, the control classes performed in a way typical of fourth graders, whereas the experimental group performed in a way typical of sixth graders' (p 28). • The experimental curriculum affected free writing. The experimental students wrote longer texts in the post-test than the control students. They were also superior in terms of Hunt's syntactic maturity indices of clauses per T-unit and words per T-unit. • The experimental students had higher scores in the Stedman Reading Structure Test. • Both white and black experimental students improved more on the rewriting tests than the control students (at 0.001 level overall and for each ethnic group separately). • Both black and white experimental students improved fluency in free writing (at the 0.025 level for black students, 0.005/0.05 for white students and 0.001/0.005 overall), but only the black students gained significantly on the measures of syntactic maturity. For clauses per T-unit, the improvement was at the 0.005 level for black students, n.s. for white students and at the 0.001 level overall. For words per T-unit, the improvement was at the 0.01/0.025 level for black students, n.s. for white students and at the 0.05/0.1 level overall. (**Note:** The first p figure had IQ as covariate, the second the pre-test.) • Only the black students showed a significant gain in reading structure (between control and experimental students) (at 0.005/0.001 levels). • As regards fluency, both groups increased the number of total words, but clause length (words per clause) did not increase significantly for either group.
Conclusions	• The experimental group gained about two grade years in terms of performing sentence embeddings in the rewriting exercise. • The omens for later syntactic maturity were good, as the experimental group developed a significantly greater willingness to reduce clauses to single words. • Grammar teaching had a useful spin-off by increasing fluency. This had also been found in Miller and Ney's (1968) study. • The intervention helped the black students in particular and would help black students in general, 'if they are like the black students in this study' (p 26). • Further research was needed to establish more effective versions of the intervention. • More appropriate measures of reading needed to be produced. • A major longitudinal test lasting several years was needed.
Weight of evidence A (trustworthiness in relation to study questions)	Medium to high • It is high because it is a well-conducted study with high reliability and validity in the data analysis. There is considerable coherence within the confines of the study itself.

	• It is medium because of the sampling, the limited reporting, the restricted free writing pre-test and the absence of a delayed post-test. • The study was on the whole very good, given the constraints. However, the constraints do affect trustworthiness. The absence of a delayed post-test is also unfortunate, as it is not clear how much long-term learning has taken place.
Weight of evidence B (appropriateness of research design and analysis)	Medium
Weight of evidence C (relevance of focus of study to review)	Medium to high In several respects the weight of evidence is high, but the lack of examination of teaching quality is a problem, given the results at the level of ethnic group. The method of allocation is also a possible problem. Much depends on the modern view of T-Units. The reading structure test scores reflect an interest at the time in variations on cloze testing, but even the researchers admit they did not understand what the scores showed. For the present study, the effect is simply to highlight that sentence combining created some spin-off effects on other language skills.
Weight of evidence D (overall weight of evidence)	Medium to high Within the limits described above, the study provides a high degree of weight of evidence, largely because it is well conducted with high validity and reliability.

MacNeill TB (1982) The effect of sentence-combining practice on the development of reading comprehension and the written syntactic skills of ninth grade students	
Country of study	Canada
Age of learners	14–15: Ninth grade
Type of study	Researcher-manipulated evaluation: randomised controlled trial (cluster)
Aims of study	• The research was designed to investigate the effect of the O'Hare 'Sentencecraft' (1975) programme on the written syntactic skill of average ability, ninth grade students, using written argumentative compositions as the data source. • A second aim was to determine the effect of the 'Sentencecraft' programme on the reading and comprehension level and speed skills of the same students.
Summary of study design, including details of sample	• The study used a true experimental design with pre- and post-test. Six classes (one each of the two high, two low and two middle ability students) were randomly allocated to experimental or control groups; therefore this was a cluster design with three clusters in each arm. • A pre-test was administered to both the groups. A nine-week treatment was conducted: the experimental class studied the Sentencecraft Program; the control group studied other areas of writing. Two argumentative compositions were assigned during and after the period. A post-test was administered after the 16-week experimental period. A delayed post-test was administered after eight weeks. • N =154 but 11 lost due to 'experimental mortality'; 75 experimental students and 68 control students (total = 143)
Data-collection instruments, including details of checks on reliability and validity	• Some explanation of instrumentation: 'The instrument for measuring written syntactic ability consisted of a single theme written on each testing occasion... ' (p 7) and the six topics for argumentative writing were listed on p 8. Topic, mode (argumentation) and time were controlled. For example, 55 minutes of the class time was given for writing each composition, with another five left for group discussion of the topic of the day. • The Davis Reading Test was used in testing reading speed and comprehension: '…it had three fairly equivalent forms and it contained a reliable measure of reading comprehension speed... Forms A, B and C were used because they were the most statistically equivalent forms' (p 9). • 'The Canadian Lorge-Thorndike Intelligence Test, Level F, Verbal Battery was administered to all students involved in the study... The Verbal Battery subtests measure word knowledge, sentence completion, verbal classification, and verbal analogies' (p 9). • Prior to the experimental-control group comparison, a check was made to determine whether the topics were roughly equivalent using the six major indices. • Validity of this method for obtaining measures of group achievement was supported by Diederich (1946) and Kincaid (1953). One-sixth of the students wrote on each occasion in order to control for its effect. Each student was randomly assigned to one of six different topic sequences. Similar arrangements for reading tests.
Methods used to analyse data, including details of checks on reliability and	ANOVA • A one-way analysis of variance was carried out for IQ scores. • A one-way analysis of the six major indices W, TU, CL, W/TU, W/Cl, Cl/TU was carried out.

validity	• A two-way analysis of variance with repeated measures was carried out to determine if there were any significant differences between the mean scores of the experimental and control groups on the pre-test and five subsequent occasions. • A chi-square test for goodness-of-fit was carried out to compare pre-test means of each sub-sample with the remainder of their groups using the major indices. • A two-way analysis of variance with repeated measures was carried out to test mean differences between reading level means across the three test occasions.
Summary of results	• No significant mean increases were found across the six writing occasions as a result of the 'Sentencecraft' programme. • Reading – level of comprehension: 'Both groups showed significant increases over the course of experiment...However, because the treatment means were not significantly different, the null hypothesis was not rejected' (p 17). • Reading – speed of comprehension: 'the experimental group thus showed significant growth between pre-test and post-test and maintained this significance over the course of the experiment despite a decline during the delay period. The null hypothesis was rejected' (p 20).
Conclusions	• Writing – As no significant mean increases were found on the six major writing indices across the six writing occasions as a result of the 'Sentencecraft' programme, the author concludes that the results of this study show that the 'Sentencecraft' programme did not result in significant increases in the mean number of selected words, phrases, or clauses that ninth grade students wrote in argumentative compositions. • Reading – 'Further, the results of this study showed that the 'Sentencecraft' program was not effective in eliciting significant growth in 'Level' of comprehension of grade nine students as measured by the Davis Reading Test (p 30). 'Nor did the experimental group show a significant mean increase in 'speed of comprehension' between pre-test and delayed post-test" (p 2).
Weight of evidence A (trustworthiness in relation to study questions)	Medium to low Insufficient data presented (writing outcomes) in order for reviewers to calculate effect sizes
Weight of evidence B (appropriateness of research design and analysis)	High RCT highly appropriate for effectiveness question
Weight of evidence C (relevance of focus of study to review)	Medium to high
Weight of evidence D (overall weight of evidence)	Medium

McAfee D (1981) Effect of sentence combining on fifth grade reading and writing achievement	
Country of study	Reviewer assumes USA
Age of learners	10–11: Fifth grade
Type of study	Researcher-manipulated evaluation: randomised controlled trial
Aims of study	The study was designed to investigate the effects of sentence-combining instruction on the reading comprehension and writing maturity of fifth grade students.
Summary of study design, including details of sample	• Randomised controlled trial • 50 fifth grade students in two grade level reading ability groups were randomly assigned to experimental and control groups • 25 in experimental group and 25 in control group
Data-collection instruments, including details of checks on reliability and validity	• The Test of Reading Comprehension (1978) • The Test of Written Language (1978) • Qualitative analysis of two free writing samples • No checks on reliability and validity other than use of published instruments and standardised instrument
Methods used to analyse data, including details of checks on reliability and validity	• Two analyses of covariance and a proportional comparison were used to determine significant differences at the 0.05 level. • Covariance data consisted of the Total Reading Battery scores on the Stanford Achievement Test (1973). • Very brief reference was given to the need to determine significant differences at the 0.05 level. • No checks were made on reliability or validity other than use of standard statistical procedure.
Summary of results	'The results showed that students who received sentence-combining instruction as defined in this study had significantly improved reading comprehension, written language, and free writing scores after treatment compared to students who received no sentence-combining instruction' (p 6).
Conclusions	'This study provides evidence that... sentence-combining instruction which includes authorship provides improvement of the use of syntax in its written form' (p 8).
Weight of evidence A (trustworthiness in relation to study questions)	Low to medium Lack of reporting of any results
Weight of Evidence B (appropriateness of research design and analysis)	High to medium RCT, but only two groups and all individuals in both groups receiving interventions together
Weight of Evidence C (relevance of focus of study to review)	Low to medium
Weight of evidence D (overall weight of evidence)	Low to medium

Mellon J (1969) Transformational sentence-combining, a method for enhancing the development of syntactic fluency in English composition	
Country of study	USA
Age of learners	12–13: Seventh grade
Type of study	Researcher-manipulated evaluation: controlled trial (cluster)
Aims of study	• To measure the effects of a novel kind of sentence-combining practice as observed in the writing of approx 250 seventh grade students over the period of one academic year • '…the main purpose of this experiment was to determine whether specially structured but a-rhetorical activities germane only to the study of grammar may yield fortuitous and quite naturalistic by-products observable in student performances in the composition class' (pp 14–15).
Summary of study design, including details of sample	• This is a cluster 3-armed controlled trial with non-random and random allocation to experimental, control and placebo groups by class and school. The students wrote before and after compositions. • N = 247 seventh grade students
Data-collection instruments, including details of checks on reliability and validity	• Nine topics in parallel A and B forms in narrative, descriptive and expository modes. • The first 10 T-units from each of the nine writing samples were grammatically analysed for 12 measures of syntactic maturity. • '…procedures were standardized during the first several weeks of work...the assistants' findings were confirmed by the experimenter for all T-units longer than twenty words, and for others on a systematic spot-check basis' (p 67).
Methods used to analyse data, including details of checks on reliability and validity	• Mean change scores on the 12 factors of syntactic fluency were analysed within each group by t-tests for correlated measures. • ANOVAs comparing mean post-test scores on the 12 factors and using pre-test measures as covariates. • Use of standard statistical procedures
Summary of results	• 'Remarkably, the experimental group experienced significant pre-post growth on all twelve factors. As anticipated, however, growth in the control group was so slight as to be virtually indiscernible' (p 74). • 'Everything considered, the experimental group as a whole clearly experienced significantly more growth than the control group' (p 87).
Conclusions	• Significant growth occurred in the writing of the experimental group • '…given the design features of the study, it would seem that this extra growth may be unequivocally attributed to the experimental treatment' (p 87). • '…a longer term experiment including a mid-test would obviously have been more convincing than the present study' (p 108).
Weight of evidence A (trustworthiness in relation to study questions)	Medium to high

Weight of evidence B (appropriateness of research design and analysis)	Medium The study is a controlled trial.
Weight of evidence C (relevance of focus of study to review)	Medium
Weight of evidence D (overall weight of evidence)	Medium

Melvin MP (1980) The Effects of Sentence Combining Instruction on Syntactic Maturity, Reading Achievement, and Language Arts Skills Achievement	
Country of study	USA
Age of learners	8–11
Type of study	Researcher-manipulated evaluation: controlled trial
Aims of study	• To study the effects of sentence-combining instruction on students' use of the conventions of written English - punctuation, capitalisation and grammar (defined as language arts achievement in this study). • To study the effects of sentence-combining instruction on syntactic maturity and reading comprehension for students aged 8–11 years.
Summary of study design, including details of sample	• Intervention study using instruction in sentence combining as the intervention for experimental group. Control group received 'normal' instruction. • N = 160. Experimental group consisted of 80 students: 20 each at ages 8, 9, 10 and 11. Boys and girls even at each age group. Same number in control group matched for age, sex and socio-economic status.
Data-collection instruments, including details of checks on reliability and validity	• Pre-tests were taken during a specific timeframe running from late November to mid-May. • Pre- and post-tests from Ohio Survey Test have nationally standardised norms. No other mention of reliability. • No details of validity
Methods used to analyse data, including details of checks on reliability and validity	• All six variables presented as mean raw scores for control and experimental groups, pre- and post-tests for each of the four age groups. • Analysis of variance for age 8 (time, time x school) and then for ages 9, 10 and 11 (time, time x school, time x school x age) with indications when significance levels reached; 'Using the SPSS computer program on pre-test and post-test scores...a univariate analysis of variance was performed to test for significance'. • No details of reliability or validity
Summary of results	'The results are mixed. There are scattered instances of significant differences between and among variables... In terms of the hypotheses, the results indicate that there is no significant difference in language arts achievement or in syntactic maturity between the two groups. Within the experimental group, there is no significant difference between boys and girls in their syntactic maturity' (p 9).
Conclusions	The results do not support any correlation between syntactic maturity or instruction in sentence combining and reading comprehension and there were no significant differences in achievement, leading to the conclusion that sentence combining is as effective an approach to instruction in language arts as traditional methods have been.
Weight of evidence A (trustworthiness in relation to study questions)	Low No indication of *what* exactly either the experimental or the control groups received by way of instruction. Concerns about selection of sample not considering level of attainment in written English, ability, aptitude, motivation, etc.
Weight of evidence B (appropriateness of research	Medium The study is a controlled trial.

design and analysis)	
Weight of evidence C (relevance of focus of study to review)	Medium It ought to be highly relevant, but seems to fall short.
Weight of evidence D (overall weight of evidence)	Medium to low Non-reporting of one of the variables and groups sizes of 20 per age level may lessen the weight of evidence.

1. Miller BD, Ney JW (1967) Oral drills and writing improvement in the fourth grade **2. Miller BD, Ney JW (1968) The effect of systematic oral exercises on the writing of fourth-grade students**	
Country of study	USA
Age of learners	9–10: Fourth grade
Type of study	Researcher-manipulated evaluation: controlled trial
Aims of study	Paper 1 • The research aims are not explicitly stated. The authors write that a 'decision was made to use modern methods of foreign language teaching...in an attempt to foster writing improvement in these [fourth grade] students'. These methods related to certain types of sentence combining using oral drilling and written exercises over a period of time. Paper 2 • The authors state: '...a classroom experiment was conducted to determine the effect of systematic oral language exercises on the writing of fourth-grade students in a typical suburban middle class school' (p 45) and to investigate 'the effect of ... (school) students' manipulation of grammatical structures on their ability to write' (p 44).
Summary of study design, including details of sample	Paper 1 • Two intact classes were divided randomly into experimental and control group. Baseline data were gathered using the IOWA Skills Test. The pre-test involved a free composition as a response to a short film. Both groups were taught traditional language arts programme. • The experimental group had the intervention in addition. They were given a six-stage intervention comprising (1) oral/choral grammar drills, (2) read text with structures and vocabulary in, (3) choral grammar drill, (4) write drill sentences (variable numbers), (5) teacher marks and (6) marks read to class. • After 1.5 months, students wrote as many sentences as they could think of, using the patterns. The 'winner' of the day was announced to the class. The post-test was the same as the pre-test. There was no delayed post-test. • N= 57: control 28, experimental 29. There were some students who took the pre-test or the post-test, but not both, and these students were excluded from the tables and the analysis; see page 98. Paper 2 • Period 1 – Two extra phases are added, as '5' and '6': (5) Final review of day's structure by T with choral repetition by students and (6) T reads sentences aloud, students write combined version. In our reported version, (5) and (6) become (6) and (7). • Period 2 – Similar to Period 1 except (a) more than one structure often practised in same session and (b) step 5 (our 6) altered at times. The review was replaced by cue and response drill for different structures. • N = 50: control 24, experimental 26 (It is unclear how far this is the same study. If it is, there is presumably a question of attrition of sample size; however, this is not mentioned anywhere.)
Data-collection instruments,	Paper 1

including details of checks on reliability and validity	• No details reported about Iowa Test • Pre/post-test: 11-minute film of fawn followed by timed writing test • Time not reported • Spellings provided on request and written on board • Unclear if brainstormed sentences ('star sentences') collected by researcher • Both groups did the same task under the same conditions. Paper 2 • Writing time limit = 30 minutes (p 45) • Film 2 (Hunter and Forest) had 'no dialogue or narration' (p 46) • Daily step 7 (our 8) exercises collected to establish if progress visible on day by day basis. Results not reported (or mentioned at all). • Some discussion about the impact of the soundtrack on stimulus films. The sound was on in Film 1 and the high incidence of single clause T-units is attributed to it. This could affect both validity and reliability.
Methods used to analyse data, including details of checks on reliability and validity	• Counting the occurrences of the structures which were practised by the experimental group (1967, p 96). • Counting 'the total number of words written by both groups on both tests' (1967, p. 97). • Hunt's (1966) measures were used: * 'the number of words per T-unit' * 'the number of words per clause' * 'the number of words in multi-clause T-units contrasted to the number of words in single clause T-units' * 'the ratio of subordinate clauses to all clauses' (1967, p 97) • Descriptive statistics • F-tests of pre/post-test score differences • ANOVA for Control/experimental group differences. No post-hoc tests were used. • No details of reliability were reported. Details of validity (Paper 1) • They began by exploring indices relating to T-units (although these were not pursued). • They also attempted to control for differential word length between the two groups. Details of validity (Paper 2) • A greater attempt was made to look for meaningful synchronic patterns, as students went through the different stages. • T-unit indices were examined in greater detail.
Summary of results	Paper 1 • The experimental group wrote more words in the post-test. • They also used more of the taught structures in the post-test, including when the different word length is taken into account. • '…little change was evident' as regards the syntactic maturity variables (= presumably non-significant)(p 97). Paper 2

	• Both groups wrote longer texts from pre-test to post-test (significant in both periods for the experimental group, 0.0001 level, but just at period 1 for the control, and at 0.01 level). Both groups wrote shorter texts from post-test 1 to pre-test 2. • The experimental group wrote significantly more taught structures in their post-tests (vs pre-tests), at 0.01 (Period 1) and 0.001 levels (Period 2). In neither case was the control group difference significant to 0.05. At Period 2 there was no measurable difference at all in terms of mean scores. Only by the end of Period 2 was the difference between them and the control group significant, but by that point the difference was significant at the 0.001 level. • The scores on post-test 2 indicated a high incidence of multi-clause sentences (significantly greater than on the pre-test, to the 0.01 level). This dropped to non-significant by Period 2. • The control group development pattern differed from that of the experimental group. The control group increased the number of single and multi-clause T groups significantly in Period 1, but there was no significant increase in Period 2. The experimental group increased both types significantly in Period 1, but by the end of Period 2 only the increase in multi-clause T-units (i.e. the taught structures) was significant. Similarly, the subordination ratio increased from post-test 1 to 2 for the experimental group, but decreased for the experimental group. No tests of significance were carried out.
Conclusions	Paper 1 (p 99) • Fourth grade students of average ability can be conditioned to produce certain structures in their writing with an increasing frequency by use of a combination of written and oral skills. • Students gain facility in writing through the use of a combination of written and oral drills, so that, within a specified amount of time, they can produce longer compositions measured in terms of the total number of words in each composition. Paper 2 (p 61) • Students who participated in these exercises wrote with greater freedom and facility than those who did not; hence, these students could write a greater number of words in less time. • Students who practised certain sentence structures in their oral and written forms used these structures more frequently than those who did not. • Students who practised putting together sentences in their oral or written form, so that simple sentences are formed into complex sentences, used a greater proportion of complex sentences. • 'For these three reasons, it has been judged that oral and written exercises have a favorable effect on the writing of fourth graders' (p 61).
Weight of evidence A (trustworthiness in relation to study questions)	Medium It is not entirely clear what were the important factors in the method or whether student (or staff) motivation was relevant. The 1968 (Paper 2) study is more convincing, although the lack of sampling data make it less trustworthy. It is also unclear whether 1967 and 1968 represent the same study, but with seven dropouts in Period 2. Again, it needs to be noted in both cases that certain uncontrolled factors detract from the degree of trustworthiness.

Weight of evidence B (appropriateness of research design and analysis)	Medium The study is a controlled trial.
Weight of evidence C (relevance of focus of study to review)	Medium The sample and context are relevant. It might be suggested that writing quality is somewhat narrowly conceived, as are the measures used to evaluate it. It would have been more helpful to have tried to establish what aspects of the method were important and to have justified the particular syntactic measures in more detail: were they simply easy to teach and test, or were adjectives and relative clauses known to cause problems in fourth grade writing? The third phenomenon, adverbial clauses, is recognised as problematic in UK and USA children's and adolescents' writing. It would have been useful if the study had provided separate indices of improvement for the different grammatical phenomena.
Weight of evidence D (overall weight of evidence)	Medium

Ney JW (1976) Sentence combining and reading	
Country of study	USA
Age of learners	10–11: Fifth grade
Type of study	Researcher-manipulated evaluation: pre- and post-test
Aims of study	• A research study looking at the effects of two modes of sentence combining instruction on writing skills (p 1) • One mode is written; the other is the mixture of oral and written similar to that used in Miller and Ney (1968).
Summary of study design, including details of sample	• This is a pre- and post-test study, with 40 students from two fifth grade classes in two experimental groups. • The two groups were taught different sentence-combining methods for 12 weeks. Each group appears to have been an intact class. • Two teachers rotated. The text is vague about the dates and periods. It is stated on page 4 that each group had different teachers weeks 1 to 4 and weeks 5 to 8. In addition, two more rotations (of how long?) were planned, but an Indian Arts teacher took one or both groups for two weeks. On page 1, the project is stated as lasting 12 weeks and three days. • The interventions comprised the following: (a) Individualised class – Four-step procedure: (1) read passage; (2) answer comprehension questions; (3) join sentences in writing; (4) break long sentences into smaller sentences. (b) Group class – Replaced step (4) with a complex procedure involving oral skills as follows: (i) read text chorally (unclear if different from Step 1); (ii) listen to teacher presenting 'concept' (? sample sentences) on blackboard; (iii) 'choral-oral' practice of sentence combining; (iv) writing combined sentences from oral cues. • The overall format was pre-test – intervention – post-test 1 – post-test 2* (* by accident rather than design). • The testing required students to write a free composition as a pre-test after viewing a short film. Three months later they wrote a post-test composition based on a different film, contrary to what was planned. The pre-test film was rescheduled for a delayed post-test a month later (but the results are not reported).
Data-collection instruments, including details of checks on reliability and validity	• Free composition followed a short film. • As film 1 ('Whose Garden Was This?') was not delivered for the post-test, a different film was substituted ('A Chairy Tale') and the original scheduled as post-test 2. The films were of unknown length and no details are provided about soundtrack. Given the extended discussion in Miller and Ney (1968) on this point, the omission is significant. • No time limit is stated, although, again, Miller and Ney (1967, 1968) did employ one. • The miscue analysis was carried out on 36 students (18 from each group) using a text from Goodman and Burke (1972). Due to recording difficulties, five students read a different text. It is not reported at what point during the study the miscue recordings took place. • There are no details of reliability, except for establishing the impact of reading proficiency. • There are no details of validity except that the order of presentation in the intervention was controlled for and references to previous work on the data-collection instruments are given.

Methods used to analyse data, including details of checks on reliability and validity	• Free writing scores were analysed by ANCOVA (unlike Miller and Ney 1967, 1968). • Miscue analysis was analysed as errors per 100 words and in terms of proportion of non-self-corrected errors which altered the meaning. • A Pearson product-moment correlation and an ANOVA test were used to check whether or not the composite miscue score was equivalent to the traditional 'miscues per hundred words'. • Miscue analysis was used to check reading level.
Summary of results	• The results are not straightforward, as different students were included in different calculations. • The difference between the scores of old and new students were not significant. • 'The greatest number of scores which proved to be significant at or below the .05 level of significance was found in a comparison of the Group and Individualized students in a comparison of the pretest using Whose Garden is This (Sept.17) with the posttest using A Chairy Tale... All but one of the scores, the scores for adverb clauses in initial position, favor the group class' (p 10). • The impact of the delayed post-test is not evaluated statistically, but the author comments that there was a fall-off by the experimental group: 'they had lost some of the abilities gained in the experimental class and thus made scores similar to those of the students in the individualized class' (p 10). • A miscue analysis was carried out as a control on the impact of the experimental methodology on reading behavior: 'In any case, the Pearson product moment correlation of the scores of the individual class and the oral group class was 0.4503 and the correlation of the miscues per hundred words of the individualized and the oral group class was 0.5573. Neither value was significant at the .01 level of confidence...No significant differences existed between the two groups' (p 14).
Conclusions	• There are no conclusions in the text or the resumé (or discussions about any conclusions). • The nearest is that the intervention favored the Group (oral) method: 'In either case, the mode of sentence combining which differed between the two classes would seem to be the crucial factor. The oral-to-written mode of the group class seemed in fact to produce superior results even though these results did not persist until the second post-test' (p 10).
Weight of evidence A (trustworthiness in relation to study questions)	Medium to low The results are what might have been expected, given the earlier Miller and Ney (1968) study. However, the lack of details about the sample and the method, the lack of clarity about the instruction, plus the absence of the more important figures (pre-test vs post-test 2), mean that the only value to the results is that the difference was in the direction expected. This is extremely unfortunate, as neither the Miller and Ney 1967 or 1968 reports give details of the individual linguistic phenomena, as this one does.
Weight of evidence B (appropriateness of research design and analysis)	Medium to low A pre- and post-test study without a control group; therefore medium to low. In some ways, the lack of a control group is not too much of a problem if one assumes that some version of the method is going to work (presumably justified by Miller and Ney 1968). The ANCOVA design is preferable to analysing separate relations separately. MANCOVA was not available at the time. The lack of baseline data is a problem and the reading scores were not inputted to the ANCOVA.

Weight of evidence C (relevance of focus of study to review)	Low The number of problems with the study and the missing information mean that, if the study is taken by itself, the weight of evidence is low.
Weight of evidence D (overall weight of evidence)	Low If the study is read in conjunction with Miller and Ney (1968), the present study can add some support to the results there. Taken on its own, however, the weight of evidence has to be low.

Nutter N, Safran SP (1983) Sentence Combining and the Learning Disabled Student	
Country of study	USA
Age of learners	6–15: First to ninth grades
Type of study	Researcher-manipulated evaluation: controlled trial
Aims of study	To examine 'the instructional efficacy and applicability of sentence combining exercises with learning disabled students' (p 2). Later expressed as being 'to explore the feasibility and effectiveness of naturalistic incorporation of SCE's into regular tutoring sessions provided for LD pupils in public schools' (p 5).
Summary of study design, including details of sample	• Quasi-experiment with an experimental group and a non-comparable control group, using post-tests to try to determine feasibility and effectiveness • Total sample is 35: with 24 in the experimental group, and 11 in the control group.
Data-collection instruments, including details of checks on reliability and validity	• '…two standardized sets of drawings were reproduced and used by the tutors as stimuli for obtaining pre- and post-writing samples...the two sets were randomly distributed for prewriting and reversed for postwriting to counterbalance any potential practice or order effects" (p 6). • No details of reliability other than the fact the two sets of drawings used for stimuli were 'standardised', although E tutors kept logs of SCE use and reactions. • Experimental group tutors designed their own sentence-combining exercises after training, but this is not necessarily a strengthening of validity. Training given to SCE E tutors.
Methods used to analyse data, including details of checks on reliability and validity	• Experimental and control groups' pre- and post-tests compared within groups on variables under investigation. Dependent T-tests were used to determine if significant within-group differences existed. Tests of significance then applied. • Dependent T-tests and tests of significance • No details of reliability other than through statistical comparison • Mention that between-group comparisons would not be valid
Summary of results	Results are basically that 'while no significant differences were present for the control group, the experimental group made significant gains on two of four measures – mean number of words and mean number of words per T-unit" (p 8).
Conclusions	The authors conclude, partly because the results support the findings of previous research that SCEs improve the syntactic complexity of varied student populations' writing, that – with some reservations – SCEs should be pursued as a feasible and effective method for improving LD pupils' writing.
Weight of evidence A (trustworthiness in relation to study questions)	Low • Small groups sizes, little known about subjects' abilities or prior exposure to SCEs. • Poor training of those delivering programme. • Slight scale of subjects' exposure to the intervention. • Wide range within that exposure. • No mention of any trialling of methods.

Weight of evidence B (appropriateness of research design and analysis)	Medium This is a quasi-experiment with a non-comparable control group.
Weight of evidence C (relevance of focus of study to review)	Medium The general focus is relevant, but the lack of contextual detail, the asymmetrical experimental and control groups, the poor conceptual focus and the indistinct sampling frame make this less than trustworthy.
Weight of evidence D (overall weight of evidence)	Medium to low Poor quality of research design and poor quality of execution.

O'Hare F (1973) Sentence combining: improving student writing without formal grammar instruction	
Country of study	USA
Age of learners	12–13: Seventh grade
Type of study	Researcher-manipulated evaluation: randomised controlled trial (cluster)
Aims of study	'...to test whether sentence-combining practice that was in no way dependent on the students' formal knowledge of transformational grammar would increase the normal rate of growth of syntactic maturity in the students' free writing in an experiment at the seventh grade level over a period of eight months' (p 35).
Summary of study design, including details of sample	• Cluster RCT • The participants were assigned as individuals to two experimental and two control classes. The students were pre- and post-tested on writing samples. • Total sample: n = 83
Data-collection instruments, including details of checks on reliability and validity	• Five written compositions in parallel forms at pre- and post-test: three modes of discourse – narration, description and exposition. • 'To enhance the reliability of their judgements, the evaluators were encouraged to read the compositions rapidly, according to the technique reported by Noyes (1963) for the College Entrance Examination Board' (p 53). • Details of validity not stated, but use of Hunt's T-units to analyse syntactic maturity of the students' compositions. The transparent presentation of the tools and the inclusion of all three writing modes (as defined in that period) suggest they were valid English composition exercises.
Methods used to analyse data, including details of checks on reliability and validity	• Six factors of syntactic maturity 1. words per T-unit 2. clauses per T-unit 3. words per clause 4. noun clauses per 100 T-units 5. adverb clauses per 100 T-units 6. adjective clauses per 100 T-units • A single qualitative judgement, based on the factors of ideas, organisation, style, vocabulary and sentence structure made concerning which of two compositions, one experimental and one control, was superior. • 'To determine whether statistically significant growth had occurred in the control and experimental groups when examined separately, mean change scores, obtained by subtracting the pre- from the post-treatment mean scores, were analysed by t-tests for correlated measures' (p 55). • Tests of significance on mean change scores of experimental and control groups • Author checked for teacher interaction – no effect. • Details of validity not applicable – means and SD of pre- and post-test scores of experimental and control groups
Summary of results	For the control group, five of the six factors of syntactic maturity showed evidence of increase, but this growth was not statistically significant. For the experimental group, highly significant growth had taken place on all six factors of syntactic maturity. The experimental group established a highly significant superiority at the 0.001 level of

	significance, over the control group on all six factors. Analysis of the data on the six factors of syntactic maturity indicated the following: • There was no evidence to indicate that the randomisation procedures had not succeeded. • The experimental group had experienced highly significant growth, at the 0.001 level, on all six factors of syntactic maturity. • The experimental group had experienced highly significant superiority, at the 0.001 level, over the control group on all six factors of syntactic maturity. • The experimental group wrote well beyond the syntactic maturity level typical of eighth graders and, on five of the six factors, their scores were similar to those of twelfth graders. • The treatment effect could not be related to the influence of a particular teacher or to whether a student was male or female. • Those with a high IQ tended to do better. Analysis of the data on the overall quality of the writing sample as judged by the eight experienced English teachers indicate the following: • The experimental group wrote compositions that were judged significantly better, at the 0.001 level, in overall quality than the control group. • Both the narrative and descriptive compositions were significantly better, at the 0.01 level, than their control counterparts. • Proportion of experimental compositions selected did not differ significantly in narrative and descriptive groups. • There was substantial agreement between the eight teachers.
Conclusions	• The author concludes that the experimental group achieved significantly more growth in syntactic maturity than did the control group. • The author also concludes 'Teachers of writing surely ought to spend more time teaching students to be better manipulators of syntax. Intensive experience with sentence combining should help enlarge a young writer's repertoire of syntactic alternatives and to supply him with practical options during the writing process' (p 76).
Weight of evidence A (trustworthiness in relation to study questions)	High
Weight of evidence B (appropriateness of research design and analysis)	High Because the experiment was designed to include two experimental and two control classes to which students were randomly assigned, the study design is highly appropriate in answering a question about effectiveness.
Weight of evidence C (relevance of focus of study to review)	High Conceptual focus, context, measures and other factors make this a highly relevant and well-conceived study.
Weight of evidence D (overall weight of evidence)	High There are high degrees of validity and reliability in the study. Combined with its appropriate focus and methods for the particular question we are trying to answer, the overall weight of evidence is high.

Pedersen EL (1978) Sentence-combining practice: training that improves student writing	
Country of study	USA
Age of learners	12–13: Seventh grade
Type of study	Researcher-manipulated evaluation: controlled trial
Aims of study	• To test the effectiveness of sentence-combining practice as a means to significantly improve syntactic and semantic fluency in student writing. • To study possible values of sentence-combining practice in helping students to more effectively conceptualise, integrate and express their ideas and feelings in writing.
Summary of study design, including details of sample	• This is a cluster controlled trial with pre- post- test and delayed post-test. • Number of participants = 113.
Data-collection instruments, including details of checks on reliability and validity	• Data-collection instruments are not stated. • There are no details of reliability or validity except that the author does say that they varied the modes of writing that the pupils engaged in, suggesting that they moved beyond narrative composition into description and exposition, in order to minimise effects associated with mode of discourse, time of writing and topic assigned.
Methods used to analyse data, including details of checks on reliability and validity	• Syntactic fluency: calculation of the number of words per T-unit judged overall quality of writing. • Matched pairs of compositions were judged holistically. • T-test of significance between post-test and delayed post-test means • Two raters were trained to check the syntactic fluency scores 1/3 of pre- and post-test and all of the delayed post-test were checked. • Details of reliability for holistic scoring were not stated. • Details of validity were not stated, although standard statistical procedures and tests were used.
Summary of results	• '…the four hypotheses investigated in this study were accepted. Seventh grade subjects trained in SC scored significantly higher than control subjects in achieving and sustaining growth in syntactic fluency...it was found that subjects trained in SC scored significantly higher than comparable control subjects on four different measures of improved conceptualisation and expression of meaning' (p 7).
Conclusions	• 'The general findings of this study, therefore, clearly suggest a strong relationship between one's linguistic ability to express ideas, feelings and experience (syntactic fluency) and one's mental ability to conceptualise and express integrated, meaningful content (semantic fluency).' • '...SC practice is a consistent, highly powerful, broadly influential tool found valuable to improve not only the HOW but also the WHAT of student writing' (p 10). • '...it must be concluded that much evidence currently calls for the widespread adoption and evaluation of SC materials in the classroom' (p 15).
Weight of evidence A (trustworthiness in	Medium to low Some controlling for confounders but threats to internal validity (no numbers in results and inappropriate analysis)

relation to study questions)	
Weight of evidence B (appropriateness of research design and analysis)	Medium Non-randomised controlled cluster trial
Weight of evidence C (relevance of focus of study to review)	Medium Conceptual focus highly relevant but lack of detail about context, sample and measures
Weight of evidence D (overall weight of evidence)	Medium Some threats to internal and external validity

Roberts CM, Boggase BA (1992) Non-intrusive grammar in writing	
Country of study	Presumed USA
Age of learners	15–16: Tenth grade
Type of study	Researcher-manipulated evaluation: pre- and post-test
Aims of study	The aim of the study is not particularly clear. One might deduce that the aim was to study whether the use of 'non-intrusive grammar instruction at the computer' would enhance students' ability to 'identify incomplete or unclear sentence structures' and their ability to identify 'sentence boundaries' (from abstract). The authors state the broad aims are: • for students to enjoy writing at the computer • for students to be able to write without initial concern for usage and spelling • to develop an awareness of the need for standard language usage • to concentrate on sentence boundary errors
Summary of study design, including details of sample	The study develops a particular intervention for an 'average' tenth grade class, based on a pedagogy of non-intrusive grammar instruction aimed at enhancing students' ability to identify and mark correctly sentence boundaries. Pre-intervention, during-intervention and post-intervention measures are used to measure the success of the intervention. N = 15 students
Data-collection instruments, including details of checks on reliability and validity	• Assumed work collected by the teacher • No details of reliability or validity
Methods used to analyse data, including details of checks on reliability and validity	• Identification of length and number of sentence-boundary errors • Simple word counts and counts of sentence-boundary errors. However, the study reports numbers of sentence-boundary errors only for the last piece of work analysed. • No details of reliability or validity
Summary of results	The authors report significant gains in fluency and a reduction in sentence-boundary errors for a number of the students in the sample. Students became more fluent, measured in terms of word count and word gain. They did not all avoid sentence boundary errors, although 12 students of the 15 'appear to be checking and then revising their sentences'.
Conclusions	The authors conclude that their findings are significant, especially the finding that '12 students appear to be checking and then revising their sentences'. They conclude that their 'results' are 'so significant (and promising)...that the collaborative experiment will continue'; and that 'voice' is heard in the writing.
Weight of evidence A (trustworthiness in relation to study questions)	Low

Weight of evidence B (appropriateness of research design and analysis)	Low: The research design is inappropriate for gauging effectiveness. There is no pre-test or secure baseline. The intervention is poorly described.
Weight of evidence C (relevance of focus of study to review)	Medium The non-intrusive approach to improving writing accuracy and quality is worthy of inclusion.
Weight of evidence D (overall weight of evidence)	Low • The sample is too small. • The study is poorly conceived and vague. • Results are conflated with, and confused with, conclusions. • The aims and objectives are not clearly delineated, and do not lead to research questions or hypotheses. • The conduct of the study is ill-disciplined and the level of analysis is low.

Rousseau MK, Poulson CL (1985) Using sentence combining to teach the use of adjectives in writing to severely behaviorally disordered students	
Country of study	USA
Age of learners	5–16
Type of study	Researcher-manipulated evaluation: pre- and post-test
Aims of study	To improve the quality of descriptive writing of behaviorally disoriented students.
Summary of study design, including details of sample	• N = 3 • Multiple-baseline across subjects • The treatments were sequential (Baseline – Treatment 1 – Treatment 2 – Treatment 3). • Descriptive praise and points were given at all stages. • All stages had a sentence-combining part and a story-writing part. • The treatments differed with respect to (a) whether the sentence-combining periods focused on the same topics as the story-writing sessions, and (b) the focus of the praise and points: punctuation, adjectives, or different adjectives.
Data-collection instruments, including details of checks on reliability and validity	• Number of words per T-unit • Number of adjectives per T-unit • Number of different words per T-unit • Comments on writing quality of stories Details of reliability • The researcher used a 17-point checklist. • Two raters were used. Details of validity: Not reported, beyond current work on sentence completion, and a commonsense approach to increasing the physical rewards given for cooperating with the study.
Methods used to analyse data, including details of checks on reliability and validity	• Plotting 'adjectives per T-unit' and 'different adjectives T-unit' against sessions and fitting a line through the plots of the baseline and each of the three treatments • Story quality was assessed by two graduate students – see instructions quoted on page 9. The person doing the assessment was asked to compare two pieces of writing (by the same student?) one written at baseline and one in a treatment phase (see page 14). • The data were graphed with a logarithmic scale to allow for the range of scores. • Presumably a simple least squares procedure was used to obtain the line of best fit. Details of reliability: Inter-observer reliability was checked, using two raters, for a sample of 25% and found to be adequately high on average for • number of adjectives per T-unit (mean = 96%) • number of different adjectives per T-unit (mean = 97%)

	• T-unit length across all conditions (mean = 93%) • story quality (mean = 94% for 18 story pairs) • the procedures checklist (mean = 100%) Details of validity: The story raters were asked to indicate what they thought they were rating for and to comment on the quality of the stories. This acted as a validity check.
Summary of results	• There was a marked increase in the number of adjectives per T-unit when Treatment 1 was introduced. The results were maintained during Treatments 2 and 3 (from 0.16, 0.36 and 0.29 to 1.14, 2.58 and 1.13 at Treatment 1 for Chad, Andy and Joe) *except* one child, 'Chad' did not receive Treatment 2. • There was a marked increase in the number of different adjectives per T-unit with Treatment 1 and number continued to rise through Treatments 2 and 3 (from 0.15, 0.31 and 0.25 to an average of 1.04, 2.87 and 1.26 for the treatment sessions). • The mean number of words per T-unit increased by 2.56 and 3.34 words across the study as a whole for Chad and Andy. This represented four and five grade levels. Chad thus went from three grade levels below to one grade level above the norm, and Andy from one grade level below to four grade levels above. Joe's grade level did not change (p 13). • The stories written during treatment(s) were judged to be better and (by one evaluator) more coherent as stories (more background information – taught – and more sequencing of actions – not part of the teaching).
Conclusions	• '…improvement in the composition skills of academically deficient students was demonstrated as a function of reinforcement and simple instructions' (p 14). This implies that the addition of sentence-combining instruction did not have an impact. • Students learned (or employed) rhetorical skills that were not being taught. In Treatment 1, adjective use was praised but not practised. In none of the treatment sessions was sequencing taught.
Weight of evidence A (trustworthiness in relation to study questions)	Medium The problem is that there is not a clear research question. The lack of discussion about validating the tests also reduces the trustworthiness slightly.
Weight of evidence B (appropriateness of research design and analysis)	Low No control group
Weight of evidence C (relevance of focus of study to review)	Medium Three factors lower this to medium: the lack of detail about validating some of the instruments used, the lack of formulated research questions and the fact that the feedback appears not to have been linguistic (thereby inevitably emphasising the import/saliency of praise at the expense of language).
Weight of evidence D (overall weight of evidence)	Medium to low

Rousseau MK (1989) Increasing the use of compound predicates in the written compositions of students with mild learning handicaps	
Country of study	USA by implication
Age of learners	9–12: three boys aged 9, 10 and 12
Type of study	Researcher-manipulated evaluation: pre- and post-test
Aims of study	To examine whether sentence combining plus reward points could improve the number of compound predicates in the free written compositions of students with 'mild learning handicaps'.
Summary of study design, including details of sample	• 'A multiple baseline across subjects experimental design was used' (p 5). • Three boys' use of compound predicates in composition was assessed over time and then specific teaching introduced. • Students received instruction and practice in punctuation and capitalisation for several 15 minute sessions, then in sentence combining. At both stages, instruction was followed by 20 minutes free writing from a picture. Each instruction and writing stage was followed by a five-minute marking and (by implication) feedback session. • No follow-up test was reported.
Data-collection instruments, including details of checks on reliability and validity	• Punctuation / capitalisation exercises, consisting of ten to fifteen sentences to rewrite • Sentence-combining task (with 10 sets of 'two or three simple sentences') • The free writing exercise required students to choose a picture. 140 pictures 'were presented to the students'. (Details are on page 5). • Reliability was established by the use of three people: the researcher, a teacher and an observer. • The researcher conducted much of the tutoring and giving instructions. • Motivation was increased by giving points and earning rewards. The number of points and rewards could be compared post hoc (and was found to be similar). • The writing samples were also scored for capitalisation and punctuation 'to ensure equal amounts of reinforcement across all experimental conditions' (p. 7). • A procedural checklist was used (and cross-checked across observer and researcher or teacher). • The free writing only has a picture (no text) as a stimulus.
Methods used to analyse data, including details of checks on reliability and validity	• Compound predicates were counted and related to T-units. • Each boy's use of compound predicates and T-units was graphed over all the sessions. • No methods of analysis were reported. • An independent measure of the outcome was done by another person; scoring agreement is reported (p 17). The figures are high (at 98%–99%), although the identification of T-Units proved harder for some participants than establishing the number of compound predicates. • 'The multiple-baseline across subjects design allowed us to determine whether the treatment procedures were effective without requiring a reversal design' (p 5).
Summary of results	• The number of compound predicates increased immediately sentence-combining reinforcement (exercises) was introduced. • The number of compound predicates per T-unit increased.

Conclusions	• Sentence-combining instruction and rewards increases the number of compound predicates per T-unit. • 'The present study demonstrated the effectiveness of the use of sentence-combining exercises and points contingent upon number of compound predicates per T-unit per 20 minutes of free writing time' (p 8).
Weight of evidence A (trustworthiness in relation to study questions)	Medium Small study
Weight of evidence B (appropriateness of research design and analysis)	Low No control group
Weight of evidence C (relevance of focus of study to review)	Low The context of the research was a specialised one and the impact of the intervention on the students may not be typical of students without behavioural and educational difficulties. The combination of sampling and generalisability problems, unexplained differential interventions and an unexplained problem with one student's scores bring the weight of evidence here down to Low.
Weight of evidence D (overall weight of evidence)	Low On balance we would opt for low because of the concerns about generalisability.

Saddler B, Graham S (forthcoming) The effects of peer-assisted sentence combining instruction on the writing performance of more and less skilled young writers	
Country of study	USA
Age of learners	9–11: The mean age of the participating students was 9 years, 3 months.
Type of study	Researcher-manipulated evaluation: randomised controlled trial
Aims of study	• The aim of the study was to examine the effectiveness of an intervention (sentence-combining instruction coupled with peer instruction) '...for improving a basic foundational writing skill, sentence construction' (p 4). The study offers two points by way of rationale: • Sentence generation is one of three major processes used in composition, hence the importance of the study focus. • More ease in generating sentences should hypothetically make available more 'cognitive' resources for other aspects of writing/composition. Hence the importance (theoretically) of making sentence generation (via sentence combining) more automatic.
Summary of study design, including details of sample	• This is an individually randomised controlled trial with two groups (experimental and control), using pre- and post-tests. • The randomisation was stratified on 'more' and 'less' skilled writing and on school, so that there was an equal number of more and less skilled writers in each treatment at each school. • The intervention (sentence combining or grammar) was delivered to the students in pairs in a laboratory-type condition. • Experimental and control groups were exposed to different interventions, only one of which (sentence-combining instruction) was predicated as having a positive effect on student writing. The total sample was 44.
Data-collection instruments, including details of checks on reliability and validity	There were three kinds of test: • Standardised tests already available to measure various literacy and oral skills • Those designed as specific to the intervention (i.e. the progress-monitoring tests) • The composition tasks which were set to gather data on writing quality, sentence combining in revision and word length Data analysis reliability is summed up as follows: • 'Students were individually tested. With the exception of the holistic quality writing measure, each assessment was scored by the first author. A second scorer who was blind to the purpose and design of the study independently rescored one third of the protocols (randomly selected). For holistic quality, two former teachers (who were blind to the purpose and design of the study) independently scored all compositions. To determine inter-rater reliability between the scores assigned by the two raters, a Pearson Product Moment correlation coefficient was calculated for each measure' (p 25). Details of validity:

	• To assess sentence combining, the researchers use an existing measure whose validity they don't question (i.e. Form A of the Sentence Combining subtest of the TOWL-3 ([Hammill and Larsen, 1996]).
Methods used to analyse data, including details of checks on reliability and validity	• Group means and standard deviations for the sentence-combining progress monitoring measure. • Both progress-monitoring measures and sentence-combining (TOWL-3) measures and quality of students' story writing measures and length of students' story measures were analysed using ANOVAs. • Reliability and validity were checked by the use of standard statistical procedures and tests.
Summary of results	Sentence combining • 'sentence combining instruction was effective in improving the target sentence combining skills' (p 29). • 'students who received sentence-combining instruction were more adept than comparison students at combining sentences following instruction....Thus, the effects of sentence-combining instruction were evident not only on the researcher-designed progress-monitoring measure, but on a norm-referenced measure of sentence combining as well' (p 31). Writing measures • 'sentence-combining instruction had a positive impact on writing quality and a number of revisions involving sentence combining' (p 31). • 'for students in the sentence combining condition, revising improved the quality of their post-test stories' (p 31). • 'students who received sentence combining instruction made more [sentence-combining] revision following instructions though the number of these was small' (p 33). • 'Following instruction, students in the experimental condition were more likely to revise their papers by combining sentences then their peers who received grammar instruction' (p 35). • In terms of writing quality, students in the experimental condition evidence a single advantage over their counterparts in the comparison condition. When students in the sentence combining condition revised their post-test papers, the overall quality of their writing improved but to a moderate degree. However, improvement in writing quality was not solely attributable to sentence combining revisions, but to other factors. Indeed, overall sentence combining instructions do not have a particularly strong impact on writing quality. Outcome measures • Average score on the five sentence-combining tests indicates statistically significant effect for sentence combining (effect size 1.31 p = 0.000). • Effect size for differences in the written sentence-combining skills of students in the treatment conditions was .86 (statistical significance at p = 0.003). • For story quality, length of story writing and sentence-combining revisions, there were no statistically significant main effects.
Conclusions	The writers conclude as follows: • 'findings from the current study replicate and extend previous research by showing that a peer-assisted sentence combining treatment can improve the sentence construction skills of more and less skilled young writers...and that such instruction can promote young students' use of sentence combining skills as they revise' (p37).

	• The study provides evidence '..that sentence combining instruction can have a positive effect on the quality of young students' writing, specifically in terms of revising the first drafts of their papers' (p. 37).
Weight of evidence A (trustworthiness in relation to study questions)	High to medium • Ethical concerns raised do not detract from the general trustworthiness of the findings. The study has been carefully designed with a large amount of attention given to issues of reliability and validity and the elimination of confounding variables (e.g. teacher effect). • The small number of participants reduces trustworthiness slightly.
Weight of evidence B (appropriateness of research design and analysis)	High Individual RCT with stratified randomisation; baseline equivalence established to eliminate chance bias.
Weight of evidence C (relevance of focus of study to review)	Medium Medium to high because of small sample size and particular learner characteristics.
Weight of evidence D (overall weight of evidence)	High to medium

Stoddard EP, Renzulli JS (1983) Improving the writing skills of talent pool students	
Country of study	USA
Age of learners	5–16
Type of study	Researcher-manipulated evaluation: randomised controlled trial
Aims of study	The purpose of the study was to examine whether or not specific training experiences in selected writing skills could result in products that achieve higher levels of quality on the variable of writing proficiency.
Summary of study design, including details of sample	This is a RCT with a complicated design: • 'The four participating districts were randomly assigned to serve as 'pull-out' districts (I and II), 'within the classroom' (District III) and a 'comparison' (District IV). The students in the first two districts and the classrooms in the third district were then randomly assigned to one of the two experimental groups. • N = 180 pupils of 'above average ability' (p 22)
Data-collection instruments, including details of checks on reliability and validity	• Holistic scoring using procedure established by the Educational Testing Service (1979) • Creativity score derived using the Moslemi method (1975) • Syntactic maturity: Hunt (1975), O'Donnell (1967), Mellon (1969) and O'Hare (1973) • The papers were each rated twice by three teams of raters. • Validity was checked by the use of procedures that had been used previously but no details are given about standardisation.
Methods used to analyse data, including details of checks on reliability and validity	ANCOVA and MANCOVA using pre-test scores as the covariate
Summary of results	• Quality of writing: Those fifth- and sixth-grade students who took part in either the sentence combining alone or the sentence combining and the creativity activities received higher holistic ratings than those in the comparison group. There were no significant differences between those who participated in the 40-minute sentence-combining sessions versus those who participated in the split 40-minute sessions (sentence combining and creativity). • Creativity of the response: Significant differences on all four measures (originality, idea production, language usage and unique style) in two comparisons – sentence combining and creativity did significantly better than sentence combining only and the comparison group. There were no significant differences between the comparison group and sentence combining only. • Syntactic maturity: Significant differences were found on all pair-wise comparisons for all four dependent measures of syntactic maturity. Sentence combining only scored significantly better on all four counts than sentence combining and creativity, and both the experimental groups scored significantly better than the comparison group.
Conclusions	• 'Sentence-combining activities can and should be introduced to above average fifth and sixth grade students'. (p 26).

	• 'Creativity activities should also be introduced to above average fifth and sixth grade students' (p 27).
Weight of evidence A (trustworthiness in relation to study questions)	Medium
Weight of evidence B (appropriateness of research design and analysis)	High to medium
Weight of evidence C (relevance of focus of study to review)	Medium
Weight of evidence D (overall weight of evidence)	Medium

Vitale MR, King FJ, Shontz DW and Huntley GM (1971) Effect of sentence-combining exercises upon several restricted written composition tasks	
Country of study	USA
Age of learners	10–11: Fifth grade
Type of study	Researcher-manipulated evaluation: randomised controlled trial
Aims of study	• '...the purpose of the present investigation was to demonstrate that language behaviors dependent upon knowledge of grammatical sentence structure were amenable to experimental manipulation' (p 521). • '…to determine the effect of several series of sentence combining exercises upon grammatically related written composition tasks under conditions in which the sentence combining drill was conducted either in an individual or in a modified observational learning paradigm' (p 521).
Summary of study design, including details of sample	• Randomised controlled trial • Two subjects assigned to each of five conditions; all 10 subjects were given writing post-test; eight subjects were given test task (post-test only). • Total sample: n = 10
Data-collection instruments, including details of checks on reliability and validity	• California Test of Mental Maturity to define the sample • Test task (sentence combining) • Writing tasks to measure aspects of the sample as findings of the study • No details of reliability or validity
Methods used to analyse data, including details of checks on reliability and validity	• A one-sample chi-square test was used to determine whether the frequency of correct test responses differed among the four treatment groups. • Two separate analyses of subjects' performance on the writing task were undertaken. • A one-sample chi-square test was employed to determine whether the subjects under the five treatment conditions differed in the total number of reduced-relative, relative and factive embeddings used in re-writing the three passages. • Analysis of variance across the three writing passages for each of the five treatment groups • Pearson r between mean words per T-unit and number of embeddings per T-unit • T-tests of significance were applied. • The dependent variable used in the second analysis was the mean words per T-unit for each of the three passages rewritten by subjects. 'Hunt (1965) has defined a T-unit as one main clause plus whatever subordinate causes are attached to it or embedded within the main clause. The W/T statistic has been shown (Hunt, 1965) to be an objective and valid definition of syntactic maturity in writing' (p 524).
Summary of results	• The experimental groups obtained significantly more correct responses than the control groups on the complex exercises. • The observers and individual learners used a greater number of words per T-unit than the active learners or the two controls on composition tasks requiring the re-writing of a short segment of prose. • A significant correlation was found between words per T-unit and the frequency with which the embeddings

	practised in the exercises were used on the composition tasks.
Conclusions	'The differences in performance on the writing task among the five treatment groups as indexed by the W/T statistic illustrated the importance of exposure to the sentence-combining exercises which composed the learning material. However, due to differences in performance between the paired observer and the individual learner groups compared to the paired learner group no general statement can be made about the reliability of the effect of exposure to the learning material independently of the conditions under which it occurred. Under the paired learner group, which on the writing task performed at the same level as the two control groups not exposed to the learning material, the paired observer and individual learner groups did use the kind of embeddings upon which they were drilled in their writing task' (p 524).
Weight of evidence A (trustworthiness in relation to study questions)	Medium Extremely small sample size and lack of reliability and validity of data-collection tools.
Weight of evidence B (appropriateness of research design and analysis)	High Randomised controlled trial
Weight of evidence C (relevance of focus of study to review)	Medium
Weight of evidence D (overall weight of evidence)	Medium