



CAMBRIDGE ASSESSMENT

**‘Aspects of Writing’:
Beyond an atomistic approach to evaluate qualities of features of
writing.**

Paper to be presented at IAEA 2008 conference, Cambridge.

**Sylvia Green
Gill Elliott
Nat Johnson**

‘Aspects of Writing’: Beyond an atomistic approach to evaluate qualities of features of writing.

Abstract

The ‘Aspects of Writing’ research is an ongoing, project being carried out by Cambridge Assessment. The study examines features of writing of 16 year olds in the UK. The study has investigated samples of writing from 1980 (GCE O level), 1993, 1994, and 2004 (all GCSE). The source of the writing samples came from English examination scripts.

The first two published studies in the series (Massey & Elliott, 1996 and Massey et al, 2005) used an entirely atomistic methodology. The atomistic approach discards some information about a student’s response to give a reduced description of student response based on more limited and discrete features.

The aspects of writing that were analysed in the original methodology were vocabulary, spelling, punctuation, sentence structure and non-standard English.

The studies informed the debate on the longitudinal comparability of grading standards in different years. It also revealed variations in aspects of writing that reflected changes in the curriculum and shifts in cultural values affecting how children wrote and what examiners valued. Not surprisingly, these studies have attracted a great deal of interest in the media and also in political circles.

The TES reported that

The report was interesting for the amount of precise information it produced about specific areas of language and also for its recognition that it is easy to oversimplify and that there are many different aspects of English that have competing values.

However, it was recognised that the original studies had not necessarily been designed with their longitudinal use in mind, and that if the studies are to continue at regular intervals into the indefinite future (as seems likely from the amount of interest they engender), then a review and revision of the project method was necessary. This paper describes the process of carrying out that review, and the experiences of piloting a revised method.

A significant change in the design was to expand the analyses from the original focus on a single sentence for each student to focus not only on a 100 word sample from each script but also to investigate text level features based on the whole text.

A history of the ‘Aspects of Writing’ (AoW) project.

In 1981 a report on writing and comprehension at different grade levels in English Language was produced by the Test Development and Research Unit (TDRU), part of the University of Cambridge Local Examinations Syndicate (UCLES). A section of this study described the accuracy of punctuation, spelling, grammar, usage and expression, the range of vocabulary, and sentence length. This was extracted from a sample of work comprising candidates’ fourth sentences from the essay question of the 1980 Midland Examining Group¹ English Language GCE² O level.

Sixteen years later, the Evaluation Service section of the Research Division of UCLES used the sentence samples from the 1980 study (which had been preserved as an appendix to the original report) in order to carry out a comparison between performance in English in 1980 with performance at GCSE in 1993 and 1994. The study was reported in 1996 (Massey & Elliott, 1996). Such a study was important because of the nature of the unique data that exam boards hold – evidence of candidate performance carried out under controlled conditions with no access to spell checking or external assistance. The data can be used to provide empirical evidence which then feeds into the ongoing commentary both specifically on standards, but also on education in general. The aim was to look at these features at the different grades awarded in the examination and to investigate how they performed as indicators of performance and progress.

The results of the 1996 study, comparing 1980 with 1993/4 (Massey & Elliott, 1996), showed a fairly dramatic decline in the correct use of most forms of punctuation, and spelling had greatly weakened. Of course it was important to note the changes in social values that had taken place over the course of that decade. The exam papers were very different, and expectations of candidates both in the classroom and in the examination hall had altered. Much of the formality of language seen in the 1980 sample sentences was no longer present, and candidates were writing dialogue and including idiosyncratic phrases much more. The study generated a great deal of media interest in the spring and summer of 1996.

In 2005, the same team of researchers from the same organisation (but re-branded to the Assessment, Research and Development division of Cambridge Assessment) updated the study using examination scripts from June 2004. The results of this were very positive, and were reported in 2005 (Massey et al, 2005). Although there was little evidence of the patterns of very formal language seen in 1980, the counts of correct use of various forms of punctuation were considerably improved on 1994, and there was enormous improvement at the lowest grades. For most of the elements considered, the profile of performance in 2004 was between that seen in 1980 and in 1993 and 1994. One exception was the percentage of words at lexical grade 5 or higher, where grades A and B very considerably outperformed their 1980 counterparts. So 2004 candidates seemed to be using more ambitious vocabulary, and there was also a trend towards a greater use of more sophisticated sentence structures. The report looked at spelling errors, the extent of which were found rather shocking in 1993/1994. At the lower grades the 2004 sample showed considerable improvement on the 1994 sample, and the research team were especially heartened to find only two instances of ‘text’ language amongst the whole sample. At the lowest grades

¹ Midland Examining Group (MEG) was a group of examining bodies comprising UCLES, the Oxford and Cambridge Board, the Southern Board, the East Midlands Board and the West Midlands Board. It was set up in 1981 and merged with OCR in 1998 (Raban, 2008).

² The examinations from which the study drew samples of writing were (i) the General Certificate of Education Ordinary (O) level, which was the secondary-level academic qualification that examination boards in the United Kingdom conferred upon students from 1951 to 1988 and, (ii) the General Certificate of Secondary Education (GCSE) which replaced it. The GCE O level was graded from A (high) to E (low) and the GCSE from A (high) to G (low) from 1988 to 1994. In 1994 an additional topmost grade - A* - was introduced to the GCSE examination.

there was a general upsurge in performance in 2004 – in the 1996 study we had to abandon some of the attempts to categorise G grade sentences, because they just didn't make enough sense to be able to do so, and that problem did not arise in 2004. In 1994 around a third of the grade G candidates did not succeed in writing a sentence which conveyed any meaning – in 2004 nearly all of them did. The final version of this report was completed at the end of September 2005, and it went on the Cambridge Assessment website (<http://www.cambridgeassessment.org.uk>) at the end of October. The entire report was published as a special issue of the journal 'Research Matters' in November 2005. The work again attracted significant media interest, and was presented at a seminar at the House of Commons and at the United Kingdom Literacy Association and British Educational Research Association conferences during 2006.

Given the longitudinal nature of the study we were able to consider the issue of how standards of performance had changed over time and this aspect of the work was of particular interest in terms of political initiatives and policies.

The atomistic method.

The atomistic approach discards some information about a student's response (style of expression, fluency and originality of ideas etc.) to give a reduced description of student response based on more limited and discrete features. In other words, the features coded are those which can be extracted from a few (or often just one) words. This in turn impacts upon the sampling frame. In the case of the Aspects of Writing Study, the sampling frame is a single sentence per candidate response.

The atomistic method was originally used for the "Aspects of Writing" studies in 1996 and 2005 because the available data from 1980 consisted of single sentences only. However, the method has certain advantages:

- In restricting the features used in the analysis to isolated elements of the writing it is possible to build up a bank of quantifiable features that can be recorded relatively quickly from a large sample of scripts.
- Within the UK system comparability over time is the focus of particular political interest. The comparison on a restricted range of features is intended to be invariant in relation to the task, thus providing a robust comparison over time. The focus of assessments changes over time (e.g. the type of task – narrative/comprehension/formal/informal etc), but it is considered that these features we define as atomistic (largely grammar, spelling and punctuation) considered in the Aspects study will/should always be important.
- With complex features of writing, an analysis either has to be narrative and descriptive or subject to a very strict coding frame. The atomistic features are selected to be largely non-judgemental and should therefore have better inter-rater reliability.

One further important aspect of this sort of comparison is that the inherent simplification within the atomistic method should promote good public understanding of the research.

The atomistic approach cannot provide the depth of knowledge that an ethnographic / qualitative approach might provide. However, it can provide a comparison between two candidate responses on a restricted range of features, where an in depth comparison may simply serve to show that they are different.

The samples and analyses were all stratified by grade. There is a possibility that this approach builds inherent trend structure into the data, i.e. we observe that candidates with lower grades perform poorly on our measures because our measures are similar to the criteria on which grades are awarded, hence we find a trend through circularity. We suggest that this is not the case for the atomistic features as the mark schemes used focus on the content of what is written, with typically only 5% of the marks allocated to grammar or spelling.

Time for change

The AoW study has been ongoing intermittently since 1994, with major reports published in 1996 and 2005, based upon writing samples taken from a GCE O level English examination in 1980 and GCSE English examinations in 1993, 1994 and 2004. The original design of the study was limited by the nature of writing samples collected in 1980. Furthermore, the first study was envisaged as a 'one-off', and the growth of this project has inevitably had an impact upon the fitness of the original method for the purposes to which it may be put in future.

At the end of the 2005 study it was clear from the interest that it generated, that an ongoing longitudinal study was likely to continue. Funding for a further, immediate phase of work during 2007-2008 was provided by the Department of Children, Schools and Families, which will include a major report on the 2007 cohort. Future years which may prove of interest include 2011 (when the cohort of students who have been subject to the UK National Literacy Strategy since foundation stage take their GCSE examinations) and 2014 (ten years from 2004, and twenty from 1994)

On a purely practical level, the composition of the research team has changed since the 2005 report – the original leader of the project has retired, and the single 'teacher judge' who identified all the grammatical errors in each sentence and provided a consistent link between the previous studies had retired from the study, and it was evident that a new judge or judges would be needed. This was outside our control, but contributed to it being a good moment to review and revise how the studies should proceed.

Finally, the considerable amount of media and political interest in this study has thrown up research questions not considered by the original study, and discussion and commentary on the study with and from other experts within the field have revealed both strengths and weaknesses in the methodology of the AoW study from 1980-2004.

It therefore became remarkably timely to carry out a review of the design of the AoW study, and particularly of the analyses used. A certain number of changes must be encompassed in future – most critically the change from the same single teacher judge to, most probably, a team of trained raters. However, that in itself brings with it the opportunity to expand upon the samples of writing, and to reconsider the analyses chosen. It also allowed the opportunity to improve and extend the design.

Reviewing the method and analyses used by the AoW study.

In order to effectively gauge the success or otherwise of different methods of data collection, and to gain an overview of the different features of writing which might be analysed, we began the process of revising the method for future AoW studies by reviewing a number of key studies which had taken place in the UK in the decade since the inception of the AoW project.

These studies were:

- NFER APU Language Monitoring/Performance Projects 1979-1988 (Gorman et al, 1988)
- Aspects of Writing Study 1 (Massey & Elliott, 1996)
- QCA Technical Accuracy Project, (Myhill, 2001)
- Writing Support Project, (Green, 2001, 2003)
- Changes in KS2 writing, (Green et al, 2003)
- QCA Analysis of Pupil Performance/Implications for Teaching & Learning (QCA, 2004-2006)
- Aspects of Writing Study 2 (Massey, Elliott & Johnson, 2005)

Two different sets of information were sought from the review. These were (i) the **methods used** in the studies, which refers chiefly to the way in which data were gathered – e.g. the nature of the sample of text,

whether a coding team was used and (ii) the **analyses undertaken**, which concerns which features of writing were evaluated – for example, counts of correct/incorrect commas and so on.

An initial desk review was undertaken, in which tables of methods used and analyses undertaken were generated for every study, detailing which features of writing were coded, how the coding took place, for example, whether it was a judgement on a rating scale, or a count of instances, and what the outcome was from the analysis.

A meeting was then convened with a number of key researchers in the field³ to discuss the advantages and disadvantages of each coding decision and to discuss the different elements and talk through the different ways of gathering the data.

The outcome of the meeting was a set of tables, detailing the advantages and disadvantages of all the methods used and analyses undertaken, as discussed in the meeting. These are shown in tables 1 and 2. These tables and the discussions which led to them, were used as the basis for selecting a number of methods and analyses which were piloted on the scripts which had been used in the 2004 study.

Selecting methods, features and analyses for future phases of the AoW study

The ultimate selection of the methods and analyses which will be carried forward into future AoW studies took place in two stages:

- i. the 2004 scripts from which the single-sentence data reported in 2005 had been collected had been retained. These were used to carry out an initial pilot of the analyses in which we were most interested. A team of six raters were recruited, and twenty five scripts from a range of grades were used for coder training and inter-rater agreement trials.
- ii. Following feedback from coders, and information from the inter-rater agreement statistics, a certain number of analyses were amended, or dropped. The revised coding sheets were then used to carry out further training and inter-rater agreement trials using a fifty script sample from scripts from June 2007.

An issue that has become important with the new method is the need for inter-rater reliability. If we are using several raters we must ensure that they are completing their codings in the same way. Since we had kept all the examination scripts used in the 2004 study, we were able to use these in 2007 to pilot the revised method and new features, using a team of six raters, all of whom were recruited because of their experience in similar studies. Five of the same team of raters then worked upon a sample of scripts from the 2007 sample.

The final selection of method and analyses is shown in Figure 1, with expansion and some commentary about inter-rater reliability (IRR) on the following pages:

³ The group comprised Dr Marian Sainsbury, NFER; Professor Debbie Myhill, University of Exeter, Andrew Watts, Cambridge Assessment, Sara Scorey, OCR, and Pauline Sutton and Sylvia Green, Cambridge Assessment.

100 word analysis

Sentence demarcation

The first group of features that we coded concerned sentence demarcation. These are all quite self-explanatory. In defining stops both question marks and exclamation marks were included with full stops to comprise a set of end-of-sentence stops. Readers familiar with students' writing may not be surprised to learn that other stops (colons and semi-colons) were rarely seen and thus a count of these other stops was maintained, but no analysis undertaken.

- Correct stops, incorrect stops, missing stops. For the AoW study this is a new feature as a correct stop always meant the end of a sample in the single-sentence method. The 100-word method allows information on this feature to be gathered.
- Comma splices. Comma splices are a common mistake seen in students' writing where the writer puts a comma between two clauses that should be two separate sentences. These could be categorised as a missing stop, or indeed an incorrect comma. Therefore to avoid double counting a common mistake comma splices were recorded once and not recorded as missing stops or incorrect commas.
- The final feature recorded concerning sentence demarcation was capitalisation. This element was not coded by the raters but will be counted by two members of the research team and differences in counts corrected through an arbitration process.

Inter-rater reliability on these features was good in both the pilot study and the 2007 data (87 to 95%), apart from initial issues with understanding of the coding frame.

These sentence demarcation features are key measures for quantitative analysis, along with being highly understandable. Additionally, they are transparent and thus promote good public understanding of the measure.

Verbs and proper nouns

- Incorrect tense
- Subject-verb agreement
- Correct capitalisation
- Missing capitalisation

There was much discussion about what grammatical features to include for word class. The study could have covered many word classes but many would have been too detailed/specialised for the AoW study and are very time consuming to count. In considering which features to include in the study it was necessary to look at what benefits they would bring.

Commas

- Correct use
- Omission
- Incorrect use

These data proved slightly more problematic with inter-rater reliability. The inter-rater reliability in the pilot study caused some concern, with counts of correct and omitted commas achieving only 50% agreement. This had increased significantly following clearer guidance in the main study, achieving 78 to 90% agreement.

This was an important feature of previous aspects studies, and it is, of course, media friendly. It is of particular public interest in the UK in the light of recent bestselling books about this aspect of grammar (Truss, 2003).

Apostrophes

The apostrophe is another important feature with regards to skills of interest to policy makers. There was consistently good inter-rater agreement of above 95% (even in the pilot).

- Possessive apostrophes correctly used
- Possessive apostrophes incorrectly used
- Possessive apostrophes omitted
- Correct abbreviation
- Incorrect abbreviation

The results for punctuation features, as expected, showed good use by higher-graded candidates and poor or absent use at lower grades.

Sentence Structure

In the 2004 study, our teacher-judge classified sentences as simple, compound, complex or multiple. It was also found that using the single sentence method the sentences from candidates below grade E often produced sentences that defied classification.

In the revised method the raters took the sentences within the 100-word sample as they should have been grammatically demarcated, rather than as they were. The fact that an error had been made was already recorded in the sentence structure rating. We also dropped the compound and complex division to give a general feel for the sophistication of the writing. However, instances of coordination and subordination were added to the codings.

- Simple sentences
- Multiple sentences
- Instances of coordination e.g. " I walked out on to the stage and in front of me was a huge crowd of people."
- Instances of subordination e.g. "All these questions were running through my mind as I sat shaking in the waiting room, where every other contestant was sat with their supporters."

Word level features

- Categorising spelling errors
- Linguistic sophistication

In the 1994 and 2005 studies the identification of spelling errors was carried out by two in-house raters who then ensured that their outcomes were the same and jointly revisited any areas of contention to ensure that a correct result was achieved. In the pilot study the inter-rater reliability for spelling errors was disappointingly poor. This may have been due to a number of factors; such as speed of work or the use of highlighter pens to record other counts, which might then have masked spelling errors. Therefore this was moved in-house for the arbitration method as used in the 2004 study.

In the past studies we had a complex method of linguistic complexity using the Hindmarsh Cambridge English Lexicon. However, we were concerned about changes in language (the Lexicon has not been updated since 1980) and had not found a suitable modern replacement. It was also a very time-consuming exercise, involving every word being looked up in the Lexicon. Therefore in the pilot we used rating on a three point scale. The inter-rater reliability was 0.87 and therefore this was retained as a useful measure.

Whole text analysis

Paragraphs

- Number of paragraphs
- Use of paragraphs

- Paragraph links

These include the counts of the number of paragraphs. Within this, the number of coherent and incoherent paragraphs are coded, together with an option for coders to note if a text is heavily dialogue laden.

Coders made a judgement about whether paragraphing was appropriately used, absent, overused or underused. Coders indicated whether or not paragraphs followed a logical order, and the effectiveness of links between paragraphs was evaluated on a simple rating scale.

Reader-writer relationship

- How crafted does the piece feel?
- How well-paced does the piece feel?
- Consistency of the narrative perspective

These were rated by the coding team on simple and straightforward scales. In the case of craftedness and pace the rating was of whether it was well crafted/paced or not. For consistency of the narrative perspective, coders decided whether the piece was lacking consistency, partially consistent or very consistent.

Discussion

The AoW study has been ongoing for more than a decade, but has begun to outgrow the original method used for the collection of data. The original studies were entirely based upon the data available from 1980, which was a sample of sentences, stratified by grade and candidate gender. The entire text had not been retained so we were limited to the data available from the single sentence. In revising the method, we decided that we would not limit ourselves to the single sentence method, even though that would mean that for some new features we would not be able to compare back with the earlier data.

We have made some major changes to the method, which we hope will sustain the project in a more robust format for the future. Key amongst these were replacing the single sentence with an 100 word sample, the introduction of counting and analysing the grammatical sentences used by candidates, and introducing whole text features, which broadens the project away from the purely atomistic approach used in the past.

We also made some changes to the features analysed, dropping some of the least successful ratings, replacing a measure of the sophistication of vocabulary based upon the Cambridge English Lexicon with a simpler rating scale, and bringing in the new non-atomistic features such as pace etc.

Certain things remain very important. The text written by the students must be written under examination conditions and must be in response to a question which requires narrative, extended writing. This has not been without its problems in the UK context, because at GCSE such writing is no longer a compulsory part of the curriculum. However, we have been able to obtain samples from an alternative to coursework paper.

One of the criticisms of the original studies was that they did not take into account the context of the whole piece. This has been brought in to the new methodology as it has been a feature of a number of studies that have been undertaken in the last 10 years. The whole text analyses are not atomistic features and could thus be vulnerable to the circularity mentioned before in our measures being similar to the criteria used in the original markscheme.

Although this methodology was carried out in the context of GCSE English it could be useful in other language teaching and learning contexts as a means of investigating grading standards and indicators of progress.

These types of analyses provide information about performance at different grade levels. This is a useful method for considering progression in students' writing and which features of writing develop at different levels of performance. Assessment results can then be used to inform curriculum developers and teachers about which features of writing need greater or less attention.

References

Gorman, T.P., White J., Brooks G., MacLure, M. and Kispal, A. (1988). Language performance in schools : a review of APU language monitoring 1979-1983. London: HMSO.

Green, S. (2001). A study of the effects of content and structural support in writing tasks. Paper presented at the 12th European Conference on Reading. Dublin, Ireland. July 1-4 2001.

Green, S. (2003). Exploring a Paradox in Children's Writing. An Investigation of Evidence which Suggests that Task Support may not have the Desired Effect. *Education* 3(13). pp.19-21.

Green, S., Johnson, M., O'Donovan, N., and Sutton, P. (2003). Changes in Key Stage Two Writing from 1995 to 2002. A paper presented at the British Educational Research Association Conference. University of Edinburgh, 11-13 September 2003.

Massey A.J. & Elliott G.L. (1996). Aspects of Writing in 16+ English Examinations between 1980 and 1994. Occasional Research Paper 1. University of Cambridge Local Examinations Syndicate.

Massey, A.J., Elliott, G.L. & Johnson, N.K. (2005). Variations in aspects of writing in 16+ English examinations between 1980 and 2004. *Research Matters: A Cambridge Assessment Publication, Special Issue*, November 2005.

Myhill, D. (2001). Better writers - applying new findings about grammar and technical accuracy. Westley, Suffolk: Courseware Publications

QCA (2004-2006) Analysis of Pupil Performance/Implications for teaching & learning
<http://www.qca.org.uk/itl.html>, accessed on 23/01/2007

Raban, S. ed. (2008) Examining the World. Cambridge University Press. Cambridge. 2008.

Truss, L. (2003) Eats, shoots and leaves. Profile Books Ltd. 2003.

Table 1: Methods used in evaluating writing.

		Disadvantages	Advantages	Other comments
*	Single sentence of text coded	No words/opportunity for error varies. Amount to be keyed varies fairly consistently by grade. Sentence not necessarily typical of entire text.	Self-contained unit of writing.	
	1st 10/20 lines of text coded	Handwriting size affects number of words/opportunity for error.	A manageable amount of writing to code, which may allow evaluation of features such as paragraphing.	
	Whole text coded	Time-consuming to gather evidence.	Allows evaluation of text based features such as cohesion and coherence, and additional features of writing (such as paragraphing) which only become significant in larger samples of writing.	The opportunity to use whole texts, and analyse a host of new features will strengthen the AoW study in the future. It is unlikely that whole texts on a GCSE examination question would ever be prohibitively long or time-consuming to code.
*	Sentence defined by candidate	Cannot use texts where the candidate has omitted full stops altogether. Sentence can be very long.	Can evaluate proportion of 'run-on' (comma spliced) sentences. Clear indication of lower-grade increase in sentence length when graphed.	
	Sentence defined grammatically	Manipulating the sample written by the candidate. In general researchers avoid such manipulation.	Controls the limitations imposed when the candidate defines the sentences.	On balance, grammatically defining the start of the sample is to be preferred.
	Narrative vs non-narrative	Candidate performance on various of the measures under consideration is likely to vary according to the type of writing elicited. For example, dialogue is far more likely to be a feature of narrative writing than it is of non-narrative. Many previous studies have used samples of both, and to limit the sample to one or the other limits the conclusions which may be reached.		It is imperative to maintain a narrative strand to the sample, since this is the only form of writing for which we have data from 1980-2004. The cost of adding a non-narrative sample of similar size to the project is likely to be too great to be viable. However it should be borne in mind that non-narrative writing samples are readily available from GCSE English examinations, and if it were possible to add a non-narrative strand at a future date it would further strengthen the study. One possibility might be to carry out a non-narrative study in one of the intervening years between narrative studies.
	Use of significance tests	Can overcomplicate simple quantitative analyses.	Can provide a measure of confidence in the results.	Previous AoW studies have deliberately eschewed tests of statistical significance. Future studies should include statistical tests as and when appropriate.

Table 1: Methods used in evaluating writing *continued*.

		Disadvantages	Advantages	Other comments
	Stratification by grade	Only useful in assessment – based research. Other studies stratify by age. Can become circular – examination candidates get lower grades with poorer performance (and vice versa). Research shows similar pattern because researchers are measuring the same features as the mark scheme which led to the grades.	Provides a useful method of sub-dividing results. Trends become readily apparent.	
	Use of examination scripts	Questions vary from year to year. Question style can vary, and it can be difficult to find sufficiently comparable stimulus material when a decade has elapsed between the scripts compared. Usually first-draft material – candidates do not generally have sufficient time to re-draft.	Confident that the work is unaided and unaffected by use of spell-checking software/dictionaries etc.	The use of examination script evidence is enormously valuable in the standards debate, as long as it is used responsibly and carefully.
	Use of purpose designed test materials	Test materials must match curriculum in use at the point in time that measurements are made, otherwise the results will be subject to bias. This renders the use of the same stimulus material (unless it is very general) difficult.	Control over what is tested. Can target specific spellings, or use the same stimulus material on different occasions over time.	
	A single judge of grammatical features	Unlikely to retain a single judge over a very extended period of time. Can be an arduous task for one person.	Where a study is small enough to enable this to occur it can be a strength. Judgements are consistent, and can remain so over time.	
	A team of raters	Need to be trained, which has cost implications. Need to establish inter-rater agreement.	Team members can be replaced if necessary without upsetting the balance of the team unduly.	

Table 2: Analyses undertaken in evaluating writing.

		Disadvantages	Advantages	Other comments
	Clause structure			
*	Number of words per sentence (average)		A straightforward quantitative measure. Easy to code using word processing software.	This formed a valuable part of the original AoW studies, and is strongly related to grade.
*	Number of characters per word (average)	May not be as closely related to lexical sophistication as some researchers would like.	Straightforward to code.	
	T units		An alternative form of grammatical structure for quantitative analysis.	Was not found particularly satisfactory in studies where used.
	Word Class			
	Number of finite verbs	Time consuming to count.	Each of these provide either qualitative or quantitative measures which can be used to evaluate detailed sentence structure.	May be too detailed/specialised for the AoW study. What additional benefits would this data bring in a study of GCSE writing?
	Number of co-ordinating devices			
	Number of subordinate clauses			
	Number of abstract nouns			
	Number of other nouns			
	Number of adjectives			
	Number of adverbs			
	Number of non lexical words			
	Judgement of impact of clause structure			
	Verb agreement			
	Tenses			
	Pronouns			
	Expression			
	Flesch reading ease	Derived for US contexts and using US grade levels. Limited application for UK based studies.	Easy to calculate from MS Word™.	Have not been found to be very effective measures in other studies.
	Flesch grade level			
	Spelling			
	Count of misspellings	The average number of misspellings by grade may hide considerable variation between individual candidates.	Generates a great deal of interest. Once lists of misspellings have been compiled, they can be categorised in many different ways at a later date.	Some studies have divided the misspellings into categories as the data were gathered. Previous AoW studies have used three broad categories (misspellings/ homophones and 'unusual'). However if all spelling errors are recorded in full, detailed classification can take place at a later date.

Table 2: Analyses undertaken in evaluating writing *continued*.

		Disadvantages	Advantages	Other comments
	Punctuation			
*	Comma – all types (correct/incorrect/ omitted)			This was an important feature of previous AoW studies.
	Commas used parenthetically		If every comma used will fit into one of the categories, could collect data at this level, and merge it to get the overall comma data which would match with previous AoW studies.	
	Commas used to demarcate clauses			
	Commas in lists			
*	Colon (correct/omitted)	Rarely seen in GCSE writing samples between 1980 and 2004.		Could be incorporated into 'other punctuation' category.
*	Semi-colon (correct/omitted)			
*	Run-on/comma splice		An important feature at GCSE, since previous AoW evidence suggests that it is strongly related to grade.	
*	Plural apostrophe		Hardly ever used in previous AoW studies. Some debate upon whether it truly exists.	To be dropped.
*	Possessive apostrophe		A topic of great interest in the public domain. Relatively straightforward to code and transparent to report because clear, well-known rules exist for their use.	
*	Abbreviative apostrophe			
	Speech marks	Complex to code, especially if correct use of both single and double speech marks were to be evaluated.	There is little evidence from existing studies of the correct/incorrect use of speech marks amongst students, and therefore such evidence would be valuable.	Not previously reported in AoW.
	Other punctuation devices		Easy to capture – can be categorised at a later date.	
	Sentence construction			
*	Type of sentence	More time consuming to code with the 100 word sample than with the single sentence method, because each candidate will provide a number of sentences.	A useful, straightforward quantitative measure.	Should be collected, to maintain link with previous AoW studies.
*	Adequacy of construction	Qualitative judgement, which depends upon many factors.		Used in AoW studies previously, but never especially successful. Could easily be replaced by text organisation data.
*	Effective communication	A 'positive' measure – looking at what students can do, rather than counting mistakes.	Becomes too complicated to judge when 100 word method is used.	Is a straightforward and effective qualitative rating for use with a single sentence method. Should not be used with 100 word method.
*	Capital letter omitted at start of sentence		A key set of quantitative measures of GCSE writing.	
*	Capital letter omitted from proper noun			
*	Capital letter used unnecessarily			
	Full stops			Not part of original AoW method because of single sentence sampling. A move to the 100 word method would provide a means of evaluating this important feature.

Table 2: Analyses undertaken in evaluating writing *continued*.

	Disadvantages	Advantages	Other comments
Non-standard English			
* Count and record of NSE	Time consuming to judge	Currently held in some importance, with the increase in text-messaging and informal speech.	Worth continuing, despite the time consuming nature of this measure.
Vocabulary			
* Lexical rating of each word (Cambridge English Lexicon)	Lexicon may become outdated. Enormously time-consuming when the sample was a single sentence. The increase in words if the 100 word sample is used would make this prohibitively expensive.	If the Lexicon upon which this is based is entirely inclusive and up-to-date, this measure goes much further than any other study in evaluating lexical sophistication in a precise, quantitative way.	In view of the rapidly aging Lexicon, and the increased burden of the 100 word sample this measure should be discontinued. However some effort should be made to see whether an electronic Lexicon is being, or could be, developed.
Judgement of range & sophistication of vocabulary.	Qualitative rating, dependent upon several factors.		A simple qualitative rating which could replace the quantitative lexical rating.
Paragraphing – ratings carried out on full text			
Number of paragraphs			Not part of original AoW method because of single sentence sampling. A move to full text analysis would provide a means of evaluating these important features.
Use of paragraphs			
Links between paragraphs – conjuncts/adverbials/other			
Structural patterning			
Textual Organisation – ratings carried out on full text			
Effective opening	Judgemental decisions which may be difficult to define/code. May be problematic if GCSE students are given the opening (as they were in 2004).	Important textual features which could replace the 'adequacy of construction' and 'effective communication' scales used in previous AoW studies.	
Effective closing			
Relationship with reader		Would work well as a catch-all judgement, using the subdivisions below as prompts.	Study 2 of the current AoW programme could pilot this as an overall judgement, using the six subheadings below as prompts.
Cohesion-coherence	Subdivisions of 'relationship with reader' which may be too detailed to define/code easily.		'Prompts' for the above.
Use of synonyms as cohesive device			
Textual organisation overall			
Signalling genre			
Initiation of narrative 'problem'			
Adaptation of form to purpose			
Ideas	Need to create descriptors on a scale and ensure reliability of coding.	Gives a broader analysis, introducing important textual features.	
Appropriacy			
Imagination			

Table 2: Analyses undertaken in evaluating writing *continued*

	Disadvantages	Advantages	Other comments
Dialogue			
Use of dialogue		Effective qualitative measure.	The complexities of categorising dialogue errors has prevented many studies from getting to grips with any detail of this feature, and AoW may be no exception. However, if the amount of dialogue seen in 100 word narrative samples of GCSE English warrants it, the issue of coding specific errors/correct usages might be revisited. In the meanwhile it is likely that texts with substantial dialogue will be removed from the sample, although this will be recorded in the method.

Figure 1: The revised list of analyses to be undertaken in future AoW studies.

An 100 word sample selected from every script. The script sample will ultimately comprise 30 boys and thirty girls at each grade at GCSE. From this the following information will be coded by the coding team:

Sentence demarcation

- Correct stops, incorrect stops, missing stops.
- Comma splices.
- Capitalisation.

Verbs and proper nouns

- Incorrect tense.
- Subject-verb agreement.
- Correct capitalisation.
- Missing capitalisation.

Commas

- Correct use.
- Omission.
- Incorrect use.

Apostrophes

- Possessive apostrophes correctly used.
- Possessive apostrophes incorrectly used.
- Possessive apostrophes omitted.
- Correct abbreviation.
- Incorrect abbreviation.

Sentence Structure

- Simple sentences.
- Multiple sentences •
- Instances of coordination e.g. " I walked out on to the stage and in front of me was a huge crowd of people."
- Instances of subordination e.g. "All these questions were running through my mind as I sat shaking in the waiting room, where every other contestant was sat with their supporters."

Word level features

- Categorising spelling errors.
- Linguistic sophistication.

Information regarding sentence demographics (average sentence length, average word length) will be extracted by the research team, using keyed transcripts of the 100 word extract and computer software.

A whole text analysis, of features seen in the entire answer to the question. The script sample will ultimately comprise 60 boys and 60 girls at each grade.

Paragraphs

- Number of paragraphs
- Use of paragraphs
- Paragraph links

Reader-writer relationship

- How crafted does the piece feel?
- How well-paced does the piece feel?
- Consistency of the narrative perspective