An evaluation of Spelling, Punctuation and Grammar assessments in GCSE

Alf Massey and Trevor Dexter

Research and Evaluation Division
University of Cambridge Local Examinations Syndicate

Paper presented at the British Educational Research Association Annual Conference at Exeter University, September 2002

Abstract

New regulations for the assessment of quality of written communication in the General Certificate of Secondary Education (GCSE) will be introduced in 2003. This paper reports on the performance of the previous style of assessment for quality of written communication, that of the spelling, punctuation and grammar (SPaG) mark

Between 1992 and 2002 5% of the total marks in many GCSE subjects has been allocated to SPaG. The evaluation explores empirical evidence from candidates' scripts from a range of Midland Examining Group GCSE examinations set between 1994 and 1996. It considers the statistical characteristics of the SPaG marks and their implications for reliability and construct validity; investiges the concurrent validity of SPaG assessments by looking at the relationships between SPaG marks and variables obtained from analysing the writing of candidates' scripts; reviews the differences in performance between male and female candidates; and estimates the impact of SPaG assessments on candidates' grades. This evaluation suggests that SPaG assessments are at least as reliable as examiners' other marks and that they are not simply determined by examiners' general perceptions of the candidates' abilities, though these have some influence. This is not to say greater reliability and validity can not be achieved. It looks as if the main determinant of SPaG marks is spelling, rather than punctuation or grammar, which may be unfortunate for boys who tend to be weaker at this aspect of writing. SPaG assessments typically fail to achieve their 5% expected weight and their effect on grade is not large.

1 Introduction

Arrangements for Assessing SPaG in GCSE examinations

The requirement to include Spelling Punctuation and Grammar (SPaG) assessments in General Certificate of Secondary Education (GCSE) examinations, taken at the end of the compulsory phase of schooling by almost all children in England, Wales and Northern Ireland, was first introduced by the School Curriculum and Assessment Authority (SCAA[1]), the regulatory authority, for the June 1992 examinations. Some details have since been revised and the current regulations (reproduced in figure 1.1) date from June 1994. Approximately five percent of the total marks available must be allocated to SPaG, according to three performance criteria.

Figure 1.1 - SCAA's regulations for assessing SPaG at GCSE (SCAA 1995)

5% of the total marks available must be allocated to SPaG according to the three performance criteria below:

Threshold performance	Candidates spell, punctuate and use the rules of grammar with reasonable accuracy; they use a limited range of specialist terms appropriately.
Intermediate performance	Candidates spell, punctuate and use the rules of grammar with considerable accuracy; they use a good range of specialist terms with facility.
High performance	Candidates spell, punctuate and use the rules of grammar with almost faultless accuracy, deploying a range of grammatical constructions; they use a wide range of specialist terms adeptly and with precision.

SPaG mark allocation

	Mark total for component			
	55-70	71-90	91-110	111-130
Maximum SPaG mark	3	4	5	6
Threshold performance	1	1	1	1-2
Intermediate performance	2	2-3	2-3	3-4
High performance	3	4	4-5	5-6

In the performance criteria vague terms (e.g. ‘limited’, ‘good’) differentiate between levels and are open to different interpretation. There are no examples within the regulations but SCAA did produce illustrative materials in a limited range of subjects; although even these were not suitable for briefing markers. Threshold marks for the award of GCSE grades must be determined via overall judgements about the attainment of candidates in the subject examinations, including assessments of SPaG.

Examinations in an increasing number of subjects have been exempted from SPaG since 1992. Subject areas with substantial or total exemptions in 1996 are shown in figure 1.2. Most exempt subjects have little extended writing, except English examinations - which include spelling, punctuation and grammar as an integral feature. However in subjects where SPaG is still assessed, coursework assessments by teachers (which were originally exempt) must now include SPaG assessments on the above terms.

[1] Replaced in 1997 by the Qualifications and Curriculum Authority.

Figure 1.2 - Subject areas with substantial or total exemption from SPaG in 1996

Art	English	Science
Design and Technology	Modern Foreign Languages	Music
Information Systems	Mathematics	

MEG's arrangements for standardising SPaG assessments

SCAA's regulations have to be implemented within the scheme of assessment for each of the GCSE examinations in a wide range of subjects set by each GCSE Examining board[1]. Examiners have to be told how to award SPaG marks. Midland Examining Group (MEG)[2] staff responsible for the administration of four of their syllabuses were asked to provide the materials used to brief examiners and/or teachers. These are summarised in table 1.1. The first approach employs exemplar material illustrating the quality of writing required for each level of SPaG marks. A second uses scripts available at the standardisation meeting (where examiners are briefed on all aspects of marking) to illustrate and discuss the required performance for a given SPaG mark. On this evidence, little support material is produced beyond the SCAA criteria themselves.

Table 1.1 - SPaG standardisation materials for four syllabuses

	SPaG standardisation materials
1	A set of exemplar material, with commentary plus SCAA's performance criteria.
2	A set of exemplar material plus SCAA's performance criteria.
3	SPaG marks for exemplar scripts are discussed at the standardisation meeting for examiners and these are subsequently used to co-ordinate marking, together with the inclusion of SCAA's performance criteria in the mark scheme.
4	SPaG marks for exemplar scripts are discussed at the standardisation meeting for examiners and these are subsequently used to co-ordinate marking, together with the inclusion of SCAA's performance criteria in the mark scheme.

Arguments for and against the assessment of SPaG in GCSE examinations

Virtues

A political consensus now recognises that accurate writing is a basic requirement for life outside school, especially in employment, and that inaccurate written work should be unacceptable in schools.

- The key aim is to raise the standard of pupils' spelling, punctuation and grammar by encouraging both pupils and their teachers to enhance these skills. Including SPaG in these high stakes GCSE assessments seeks to force teachers in a range of subjects to acknowledge that quality of writing is important.

- Arguably, candidates should be required to demonstrate competency in spelling, punctuation and grammar in the context of each subject. SPaG may in part at least be assessed in a subject specific sense, emphasising the appropriate vocabulary.

[1] In June 1996 these were the Midland Examining Group (MEG); the Northern Examinations and Assessment Board (NEAB); the Northern Ireland Schools Examination and Assessment Council (NISEAC); the Southern Examining Group (SEG); the University of London Examinations and Assessment Council (ULEAC); and the Welsh Joint Education Committee (WJEC).

[2] The Midland Examining Group now forms part of Oxford, Cambridge and RSA Examinations

- More arguably, SPaG assessment might also improve the marking of subject content if awarding marks explicitly for SPaG provides a vehicle for examiners to express their exasperation at poor writing, or vice versa. It could discourage examiners from consciously or unconsciously deducting/awarding marks for writing without the sanction of the mark scheme, although this problem may be minimal in well managed assessments like UK public examinations. For instance Massey (1983) studied the effects of penmanship and complexity and accuracy of prose, amongst other incidental variables, on achievement in GCE A level English: reporting that they were unrelated to the marks awarded.

Concerns

Arguments against assessing SPaG in GCSE do not deny the importance of writing quality. They stress the inherent difficulties of using SPaG assessments to motivate pupils and teachers. There are several areas of concern:

- The main issue of principle asks if it is valid to test writing skills as part of the examination of an academic subject. SPaG marks will change the rank order of candidates. Should not the rank order of candidates in a subject be determined exclusively by their performance in the subject? Candidates who know nothing can obtain 5% of the available marks from good language.

- Discrimination becomes an issue. Assessing SPaG may be hard on those candidates whose mother tongue is not English and may also disadvantage boys by exacerbating the effects of sex differences in language skills. For instance Massey & Elliott (1996) studied changes in standards in writing in 16+ English examinations between 1980 and 1993/4 and noted that boys tended to use a slightly richer vocabulary and marginally more ambitious grammatical structures. Girls made fewer spelling errors and were less likely to make some types of punctuation error than boys, but were equally likely to have used their chosen grammatical forms correctly. Such variations might attract differential rewards/penalties in SpaG assessments

- SPaG performance makes it more difficult for examiners to set accurate grade thresholds. The scripts the examiners consider to decide if a given mark deserves a given grade will have different mixtures of marks for subject knowledge and SPaG. Balancing sound subject performance and poor SPaG and vice versa complicates these judgements. The size of the examiners' 'zone of uncertainty' for their judgements will probably increase. An allied minor problem is that where SPaG marks are added to an existing mark scheme, components often have a maximum mark which can prove difficult for examiners to work with: it is somehow harder to visualise 20 out of 63 than 20 out of 60. Variability in SPaG also complicates the selection of archive scripts to exemplify grading standards.

- There are numerous reliability and marking issues. The subjective nature of the examiners' judgements concerning the broadly phrased SCAA SPaG performance criteria are central. There may be little or no consensus and examiners are often openly unhappy about the need to make these assessments. And how competent are examiners in subjects other than English to mark SPaG? These are not easy questions to resolve and unreliable SPaG assessments will reduce the reliability of the examination. For instance:

1. Are candidates who write longer pieces of work or those who try to use more difficult language penalised?
2. Can performance in the areas of spelling, punctuation and grammar be combined into a single marking judgement? How should a candidate be assessed if s/he has an uneven profile of spelling, punctuation and grammar? Are all three concepts equally important?
3. Are certain types of error more important than others (for instance those involving a discipline's specialist vocabulary)?
4. How should examiners treat scripts where SPaG worsens towards the end of the exam, perhaps under pressure of time, or scripts where a candidate makes few, but fundamental, errors throughout the entire script?
5. SPaG marks might be influenced by the overall academic quality of a candidate's script (i.e. a halo effect) or by the quality of the candidate's handwriting. Can we guard against these effectively?
6. How too are we to ensure all examiners utilise the full SPaG mark range (and do so equally) instead of being conservative and tending towards a middle mark?

- On a very practical level SPaG might make marking scripts more time consuming. It certainly discourages examiners from marking a question at a time, compared to a candidate at a time, which is many ways a strategy to be encouraged. Solving some of the questions above via more sophisticated SPaG assessments might be resisted if they take more of examiners' time (and perhaps in consequence cost more).

- There are comparability issues here too. Are SPaG marks equivalent across tiers and is it important that they should be? Is assessment of SPaG consistent across subjects and/or examining boards and does this matter?

Improving spelling, punctuation and grammar is clearly a good thing. But do the benefits of SPaG assessments in GCSE outweigh such drawbacks? A considered evaluation is perhaps overdue; not least because to say that assessing SPaG has never been popular is a considerable understatement. For instance one Chairman of Examiners for Religious Education (Owen, 1992) claimed that 'the degree of subjectivity and of chance as to what mark a candidate received, would be wholly unacceptable to anyone with a concern for fairness to all and for the professional integrity of examiners'. He saw SPaG assessments as 'no more than a political ploy to appear to be raising educational standards' and complained that 'candidates are being penalised for ..... skills which have often never been emphasised in previous years of their education'. He saw the specific reward of English skills in examinations assessing history, geography, biology etc. as 'bizarre' and even suggested that many of the examiners might lack such skills themselves; doubting their competence to assess them. He laid the blame at the politicians' door - 'Professional educators - teachers, examiners, examination boards and SEAC[1] - have opposed the blanket introduction of SPG marks on the grounds of impracticability, unfairness and inappropriateness - but their professional expertise and judgement have been ignored. The former Secretary of State for Education and his colleagues are entirely to blame for this travesty of educational justice'. Like many others at the time this examiner saw SPaG assessments as a stick, the purpose of which was to either to beat the teaching profession into emphasising basic language skills, or to lever up educational standards, according to one's point of view.

[1] SCAA's predecessor as the regulatory body.

Previous research into SpaG assessments

Following the introduction of SPaG assessments into GCSE in 1992, the GCSE examination groups conducted a joint programme of research into the initial SPaG assessments. Adams (1993a) reported some of the statistical characteristics of the 1992 SPaG marks, noting many differences between subjects and between groups in average marks awarded. There were variations relating to types of questions: long essay examination papers (i.e. English Literature) had lower SPaG marks than question papers requiring little extended writing. But mean SPaG marks were on the whole quite high, ranging from 52% to 95% in papers testing the full GCSE range of candidates. SPaG was moderately correlated with attainment (most correlations for papers involving the full ability range falling in the range 0.4 - 0.6). Between component SPaG correlations were lower (most ranging from 0.2 - 0.4) than those between SPaG marks and subject marks and correlations between SPaG marks awarded via different subject were generally lower still (mostly 0.1 - 0.3, although values as high as 0.51 were reported). Limited evidence suggested that females slightly outperformed males of equivalent subject ability on SPaG.

The tone of this inter-group report was negative, reflecting 'popular' opinion of the day. For instance it concludes that 'there may be variation in the application of the criteria for SPaG across groups and across subjects within groups'; that 'correlation evidence suggests either that SPaG is not a clearly identifiable trait in individual candidates, but rather a nebulous and ill-defined concept, or else that there is a great deal of unreliability in its marking, or both. In any case, the effect on candidates' grades was small'.

This last conclusion seems surprising when the changes introduced by SPaG marks were estimated to have had an effect on the grades obtained by between 5% and 16% of all candidates; the percentages varying across the range of syllabuses observed. In some cases the effects were even-handed, with as many candidates receiving improved grades as worse ones. But in several examinations the numbers obtaining better grades as a result of the introduction of SPaG assessments were substantial. This effect was largely a product of two factors. In 1992, in MEG at least, grade thresholds were set in the (previously) 'normal' way, with the SPaG marks which had been hastily tacked on to existing marking schemes excluded from consideration. The grade thresholds chosen were then increased to allow for SPaG marks - by the minimum mark for the top SPaG level at grade A; by the minimum mark for the middle level at grade C and by the minimum mark for the bottom level at grade G. However, where large numbers of candidates were awarded relatively high SPaG marks, as appeared widespread in 1992, many candidates with total marks for subject content just below the 'original' grade thresholds would have obtained more than the minimum additional marks required and consequently achieved a higher grade. It seems quite possible that this minor feature of the then recently introduced Mandatory Code of Practice for the conduct of the GCSE was one of the sources of the widespread improvements in GCSE grades at this time. Although negative the report pointed out that these were early days for SPaG assessments and anticipated improving standardisation with future training and experience.

A Midland Examining Group contribution (MEG 1992) to the Inter-Group Research Committee programme of work reported by Adams, showed that correlations between SPaG and subject marks were higher in History and Religious Studies than Science: as might be expected given their relative dependence on writing skills. The range of SPaG marks was larger in Science than History and Religious Studies, with science examiners thus appearing more willing to differentiate in this regard. A WJEC report on SPaG

assessments in the 1993 GCSE examinations (Adams 1993b) showed SPaG marks very highly correlated (0.87) with subject content marks in a GCSE English examination, but less well correlated with content in English Literature (0.61) and History (0.39 - 0.6). SPaG is of course an explicit assessment criterion within English. The correlations between SPaG marks across subjects (ranging from 0.33 to 0.5) were lower than the correlations between corresponding correlations involving marks for subject knowledge (which ranged from 0.59 to 0.69) but the former would be reduced by the restricted variance and inherent unreliability of the brief SPaG assessments and there was no attempt to estimate the underlying levels of correlation for either SPaG or subject knowledge across subjects. Adams here concludes 'that three complex concepts ..... can be apprehended in a single mark, based on flimsy descriptions of performance, is perhaps optimistic', arguing that examiners' assessments of SPaG may be 'mostly governed by their view of the candidates' subject performance'.

An outline of this evaluation

This paper reports an attempt to evaluate the current procedures for SPaG assessments in GCSE, now that some years have elapsed in which the regulatory body and examining boards have had the opportunity to refine their approaches. It may be timely, since it seems likely that arrangements for assessing the quality of writing may be revised when new GCSE syllabuses are introduced for examination in 2003. The work illustrates the effects of the current arrangements by exploring empirical evidence from candidates' scripts from a range of MEG examinations (set between 1994 and 1996), largely collected as part of the examining board's ongoing program of examination evaluations. Details of SPaG assessments and marks for subject content were available; these having been extracted clerically from random samples of scripts. In addition, information has been gathered from samples of candidates' writing taken from scripts from one examination, which have been analysed with respect to features of spelling, punctuation and grammar and other potentially relevant variables. These features may supplement the information from examiners' marks and help to shed light on the reliability and validity of the SPaG assessments and their impact on grading.

- Statistical characteristics of SPaG marks are surveyed to provide an overview of SPaG assessments in operation. Frequency distributions, means, standard deviations, correlations and achieved weights of SPaG marks are reported and the implications of these statistical characteristics for the reliability and construct validity of the SPaG assessments are discussed. The limited evidence available relating to the stability of SPaG assessments over time is also reviewed.

- Concurrent validity of some SPaG assessments is investigated by looking at the relationships between the SPaG mark and additional relevant variables (such as the proportion of spelling errors or quality of handwriting) obtained by the direct analysis of writing samples from candidates' scripts.

- Empirical evidence regarding differences between the performance of male and female candidates in SPaG assessments is also reviewed.

- The impact of SPaG assessments on candidates' grades is considered by estimating how many candidates' grades are changed as a result of their obtaining untypically low or high SPaG marks. The equivalence of SPaG assessments in different tiers/options is also considered and the importance of any lack of equivalence in relation to the equivalence of grading standards across tiers/options

is discussed

2 The statistical characteristics of SPaG assessments

Distributional information
Data from 10 examinations were available. Table 2.1 lists them, showing subjects and years, the range of components involved, mark allocations, target grade ranges and sample sizes. Summary statistics are also included (SPaG mark distributions, percentage means and standard deviations for component SPaG and subject content totals).

Where examinations are not 'tiered' and thus include candidates spanning the full range of ability, most candidates were awarded marks in the middle of the ranges available. Proportions of candidates awarded extreme SPaG marks varied noticeably between different components and syllabuses. Such variations may not be entirely random. For instance, 1996 Business Studies 1351 coursework marked by candidates own teachers included a substantial proportion of candidates gaining the maximum mark (four) available, although they were set in the context of a comparatively widely spread distribution of SPaG marks; as revealed by scrutiny of the component percentage standard deviations. Teachers awarded an equally wide spread of SPaG marks for the other coursework component (1996 History 1607) included in this study. It is possible that some students are more likely to produce sound writing in coursework than under examination conditions, so SPaG marks awarded to coursework represent genuine variations in the quality of language compared to that exhibited in examinations: although perhaps teachers are more likely than external examiners to award high or low SPaG marks.

SPaG distributions for tiered syllabuses reflect the abilities of the candidates entering each tier. Higher tier papers have more high SPaG marks and lower tiers have more candidates on lower mark points, as might be expected if examiners all work to similar marking instructions, derived from SCAA's criteria. 1996 Geography 1588 serves as a good example. Papers 1,2 and 3 are vertically differentiated alternatives and have mean SPaG marks of 19%, 43% and 87% (respectively) of the maximum available. On the (tenable) assumption that subject ability and language skills are correlated, this pattern is in itself evidence for the validity of SPaG marking, in that examiners are awarding higher marks to abler candidates, on average. Although determining whether or not examiners marking different papers are awarding equivalent SPaG marks to candidates of similar ability needs rather more investigation.

Examiners in some subjects/components (for instance History 1607 paper 1 in both 1994 (0.8% only) and 1996 (0.7%); Business Studies 1513 (1.5%)) seem loath to condemn candidates to a zero SPaG mark. English Literature 1512 examiners seemed particularly unwilling to awarded the lowest available SPaG marks; very few indeed of their candidates, even for the lowest tier papers, were awarded zero marks. However examiners in some other subjects were less tender minded: especially Physical Education, where examiners were remarkably miserly and awarded more low marks than high ones, and Drama, where SPaG marks were extremely widely distributed at both ends of the scale. Might such variations suggest that SPaG markers' judgements perhaps reflect an absolute scale of values - as the universal imposition of the SCAA criteria might indicate they should? If so, any variations in the overall quality of candidates opting for different subjects might result in variations in SPaG marks like those observed. The noticeable variations in the distribution of SPaG marks for components in 1996 Religious Studies 1730 also supports a similar 'self-selection' hypothesis, given what we know of the

Table 2.1 - Mark distributions for SPaG

Year and syllabus	component	grade range	n	% awarded each SPaG mark						SPaG		Total excluding SPaG	
				0	1	2	3	4		mean proportion	standard deviation proportion	mean proportion	standard deviation proportion
1994 History 1607	paper 1	G-A*	359	0.8	10.9	34.8	35.9	17.0		.65	.24	.48	.22
				0	1	2	3						
	paper 2	G-A*	359	2.5	11.1	65.2	21.2			.68	.22	.51	.13
				0	1	2	3	4					
1995 English Literature 1512	paper 1 (B)	G-D	398	1.3	29.6	48.7	17.8	2.5		.48	.20	.34	.12
	paper 2 (A)	F-A*	409	0.2	4.2	23.5	43.8	28.1		.74	.21	.66	.17
	paper 2 (B)	F-A*	764	0.0	6.0	29.3	41.4	23.3		.70	.21	.59	.16
				0	1	2	3	4					
1995 English Literature 1513	paper 1	G-D	686	2.5	24.3	49.7	21.4	2.0		.49	.20	.31	.13
	paper 2	F-A*	732	0.3	4.4	26.4	42.9	26.1		.73	.21	.52	.17
				0	1	2	3	4	5				
1995 Geography 1576	paper 1	G-C	338	2.1	15.4	40.8	39.1	2.7	0.0	.45	.16	.52	.15
	paper 2	E-A*	332	0.3	0.3	2.1	19.9	47.0	30.4	.81	.16	.58	.14
				0	1	2	3						
1995 Physical Education 2367	paper 1	G-A*	692	17.5	51.6	28.3	2.5			.39	.24	.42	.20
	paper 2	G-A*	692	16.8	51.6	28.8	2.9			.39	.24	.51	.19
				0	1	2	3	4					
1996 Business Studies 1351	coursework	G-A*	1761	1.5	13.6	29.6	33.3	22.0		.65	.26	.54	.22
				0	1	2	3	4					
1996 Geography 1588	paper 1	D-G	637	31.9	61.1	7.1	0.0	0.0		.19	.14	.42	.15
	paper 2	E-B	597	7.4	25.0	54.8	12.9	0.0		.43	.20	.52	.14
	paper 3	D-A*	640	0.0	0.5	5.9	39.4	54.2		.87	.16	.53	.13
				0	1	2	3						
	paper 4	G-C	618	18.6	50.5	28.3	2.6			.38	.25	.51	.17
	paper 5	E-A*	586	1.2	8.9	33.3	56.7			.82	.24	.54	.16
				0	1	2	3	4					
1996 History 1607	paper 1	G-A*	1634	0.7	10.2	33.5	41.6	14.0		.65	.22	.49	.21
				0	1	2	3						
	paper 2	G-A*	1634	4.6	21.7	56.5	17.1			.62	.25	.56	.15
				0	1	2	3	4	5				
	coursework	G-A*	1314	4.3	6.3	16.5	29.7	28.8	14.3	.63	.26	.66	.18
				0	1	2	3						
1996 Religious Studies 1730	paper 1	G-A*	605	4.1	27.6	49.1	19.2			.61	.26	.50	.18
	paper 2	G-A*	614	2.6	22.5	65.6	9.3			.61	.21	.44	.17
	paper 3	G-A*	614	11.1	27.7	43.8	17.4			.56	.30	.52	.20
	paper 4	G-A*	513	3.9	18.5	58.1	19.5			.64	.24	.57	.20
	paper 5	G-A*	608	17.4	30.6	37.2	14.6			.50	.32	.49	.21
	paper 6	G-A*	619	0.2	14.7	41.7	43.5			.76	.24	.57	.20
	paper 7	G-A*	588	8.7	23.1	50.5	17.7			.59	.28	.49	.23
				0	1	2	3						
1996 Drama 2325	paper 1	G-A*	589	8.7	30.4	45.7	15.3			.56	.28	.48	.18
	paper 2	G-A*	583	12.3	32.2	38.9	16.5			.53	.30	.51	.20

candidates for the alternative components in this syllabus. SPaG marks for component 5 (Islam) are the most negatively skewed whilst those for component 6 (Judaism) are more positively skewed than others, matching the grades awarded in these options.

For most non-tiered syllabuses the mean marks for SPaG are in the range of 53%-68% of maximum marks. The exceptions are 1995 Physical Education 2367, with a mean of 39% for both papers and 1996 Religious Studies 1730 paper 6 with a mean of 76%. Possible self-selection by candidates for both of these has been mooted above. In general SPaG mean % marks are slightly higher than those for subject knowledge in the same examination, although 1995 Physical Education 2367 and 1996 History 1607 coursework prove exceptions to this. For tiered syllabuses, SPaG mark means followed the tiering patterns, with higher tiers having higher means. The standard deviations of SPaG marks (again expressed as percentages of maximum marks) suggest that they were normally more widely spread than marks for subject content. In only one of the 29 components

studied did this fail to prove the case. Why examiners also frequently fail to use the full range available for marks for subject content falls beyond the scope of this study. Percentage standard deviations of SPaG marks in the non-tiered syllabuses were all in the 21% - 30% range. Standard deviations for SPaG marks in tiered question papers are usually smaller, again properly reflecting selective entry.

Correlational evidence regarding the validity and reliability of SPaG marks
Correlations between SPaG marks and marks for subject content in the same component (in table 2.2) range from 0.30 (1996 Geography 1588 paper 3) to 0.75 (1996 Drama 2325 paper 2). These are, broadly, in the range observed in the initial round of SPaG assessments (Adams, 1993a). Correlations between SPaG marks in different components in the same examination range between 0.41 (1996 History 1607, papers 1 and 2) and 0.52 (1996 History 1607, paper 1 and coursework). These are higher than those reported from the first year of SPaG assessments but similar to the levels of correlation reported by Adams (1993b) from a GCSE History examination set in 1993. What do these and the other correlations available mean? Interpreting such correlations is notoriously difficult, as the values obtained are affected by the reliability and the variance of the two measures correlated as well as by the strength of their underlying relationship. What do we expect? Positive correlations between SPaG marks on different papers marked by different examiners, and perhaps even in different subjects, might reasonably be seen as evidence that they are reliable 'repeated assessments' of the same trait(s). But how high a correlation might we expect? SPaG marks are a single overall rating using a maximum of only three to five marks. Where measurements are as limited as this, correlations seem likely to have a low ceiling. What of correlations between SPaG marks and subject content? We should be looking for both convergent and discriminant evidence of validity: so are correlations with other traits, notably subject knowledge, lower than those with other SPaG assessments when we might reasonably expect them to be so? Or are there indications of halo effects, whereby examiners are influenced by subject knowledge in awarding SPaG marks?

The extent of the correlational evidence available varies. These data were for the most part not originally obtained for this project and for many examinations different samples of candidates were drawn for each component, so that SPaG marks could only be correlated with subject content marks for the same component and hence awarded by the same examiner. We will consider this source of evidence first.

- Some examinations cater for the full GCSE ability range in all papers. In the 1994 History 1607 examination, correlations between SPaG and subject content marks were 0.69 for paper 1 and 0.52 for paper 2. In 1996 the equivalent correlations were 0.7 and 0.53. In 1996 Religious Studies 1730 such correlations are of a similar order of magnitude, ranging between 0.54 and 0.64 for papers 1 to 7. Here the highest correlations seem to be associated with relatively high variances in either the SPaG marks or subject content marks or both. In 1995 Physical Education 2367 (0.7 in paper 1 and 0.64 in paper 2) and 1996 Drama 2325 (0.56 and 0.75 for papers 1 & 2 respectively) correlations between SPaG and subject content marks were again of a similar order. Again there is evidence that high mark (SPaG or content) variances account for some of the differences observed. Correlations between SPaG and subject content of around 0.6 would therefore seem typical where selective entry is not an issue.

Table 2.2 - Product-moment correlations between SPaG marks and with marks for subject content

1994 History 1607 (n=359)

	p1_SPaG	p2_SPaG	paper 1
p1_SPaG	-		
p2_SPaG	.50	-	
paper 1	.69	.41	-
paper 2	.60	.52	.76

1995 English Literature 1512

	paper 1 (B)	paper 2 (A)	paper 2 (B)
SPaG mark for paper	.45	.62	.60
n	398	409	764

1995 English Literature 1513

	paper 1	paper 2
SPaG mark for paper	.55	.63
n	686	732

1995 Geography 1576

	paper 1	paper 2
SPaG mark for paper	.59	.42
n	338	332

1995 Physical Education 2367

	p2_SPaG	paper 1	paper 2	GCSE English	GCSE Maths
p1_SPaG	0.47 (691)	.70 (691)	.57 (691)	.58 (572)	.55 (571)
p2_SPaG	-	.58 (692)	.64 (691)	.54 (572)	.50 (571)
paper 1		-	.69 (691)	.62 (572)	.73 (571)
paper 2			-	.62 (572)	.64 (572)
GCSE English				-	.66 (565)

1996 Business Studies 1351

	coursework
SPaG mark for paper	.66
n	1761

1996 Geography 1588

	paper 1	paper 2	paper 3	paper 4	paper 5
SPaG mark for paper	.42	.42	.30	.40	.41
n	637	597	640	618	586

1996 History 1607

	p2_SPaG	cw_SPaG	paper 1	paper 2	coursework	business studies coursework SPaG
p1_SPaG	.41 (1634)	.52 (1314)	.70 (1634)	.61 (1634)	.54 (1314)	.36 (493)
p2_SPaG	-	.51 (1314)	.51 (1634)	.53 (1634)	.50 (1314)	.31 (493)
cw_SPaG		-	.69	.66 (1314)	.68 (1314)	.43 (379)
paper 1			-	.81 (1634)	.77 (1314)	.47 (493)
paper 2				-	.74 (1314)	.42 (493)
coursework					-	.43 (379)

1996 Religious Studies 1730

	paper 1	paper 2	paper 3	paper 4	paper 5	paper 6	paper 7
SPaG mark for paper	.55	.54	.64	.64	.59	.64	.59
n	605	614	614	513	608	619	588

1996 Drama 2325

	paper 1	paper 2
SPaG mark for paper	.56	.75
n	338	332

Coursework assessments also normally span the full ability range and in both examples present here levels of correlation between SPaG and subject content marks (both awarded by candidates own teachers) are again relatively high (0.68 for 1996 History 1607 and 0.66 for 1996 Business Studies 1351). This may reflect the opportunity for diligent candidates to earn marks for quality of presentation under coursework conditions. But given the teachers' role, the possibility of halo effects must be borne in mind.

Let us now consider tiered examinations, where alternative papers are set for pupils of different levels of ability.

- In the case of 1995 English Literature 1512, correlations between SPaG and subject content totals for the two papers (2&3) targeted at grades A*-F are 0.6 and 0.62 respectively, whereas that for the paper (1) targeted at the more restricted grade range D-G is only 0.45. Whilst mean SPaG marks for paper 1 are much lower than those for the other papers, SPaG marks in all three have similar variance. This is not the case for subject content, for which marks in paper 1 are less well dispersed, which may contribute to the lower correlation observed. It is however quite possible that the underlying correlation is weaker in the group 'selected' for paper 1. English Literature is often a curricular adjunct to English Language, taught in the same classes, and this group might include substantial numbers of poorly motivated students who have sound language skills but have made little effort to master the set books and have consequently been assigned to the lower tier.

- 1995 English Literature 1513 shows a not dissimilar pattern and level of correlation, with upper and lower tier papers exhibiting very similar SPaG and subject content mark means and standard deviations to those described above for syllabus 1512, but a lower correlation between SPaG and content marks in the lower tier than in the higher tier.

- In 1995 Geography 1576 the two papers are again each targeted at a different range of grades and although paper 1 and 2 SPaG mark totals have similar standard deviations they have very different means, reflecting their candidatures' abilities. The correlation between SPaG and subject content marks is 0.59 for paper 1 (which is thus of the same order as those often found in papers assessing the full ability range) but only 0.42 for paper 2. The standard deviation of SPaG marks for both papers is relatively low (16% of the available mark range in both cases) and subject content mark standard deviations are also only moderate, but there is nothing to explain the difference between papers in the relationship between SPaG and content.

- In 1996 Geography 1588 candidates selected one from Papers 1,2 and 3, according to ability. SPaG/content correlations in these relatively tightly targeted papers were only 0.42, 0.42 and 0.3 respectively. SPaG marks on papers 1 and 3 are rather more tightly bunched than in other components (and, typically, other subjects) and it seems likely that this may have contributed to these relatively low correlations. Candidates entering 1588 must also choose between vertically differentiated optional papers 4 and 5 and SPaG/content correlations for these are 0.4 and 0.41 respectively.

- Thus correlations between SPaG marks and subject content totals are generally low here. Where papers are targeted at a restricted range of ability the relationship observed between SPaG and subject content marks seems likely to be weaker, with correlation values of around 0.4 being more typical in these circumstances.

In a few instances we have matched data across the different components in examinations (or even across different examinations) available. These give us more scope. Correlations between SPaG marks in different components (and thus awarded by different examiners) may be evidence of repeated measurements of the same trait, bearing upon reliability. Patterns of correlations between SPaG marks and subject content in the same and different components may be considered to see how they match our expectations and thus begin to provide evidence relating to both 'reliability' and 'construct validity'. Data of this type are available for both the 1994 and 1996 examinations for History 1607 and for 1995 Physical Education 2367.

- Correlations between SPaG assessments from different components are, mostly, in the region of 0.5. Given the very brief assessments involved this does not seem unreasonable. We cannot expect measures which are so short to produce correlations between repeated measurements of the order of 0.8+ normally sought in estimates of the reliability of entire tests. The adequacy of the reliability of SPaG assessments will be investigated further below, using question level data.

- In the one set of cross subject SPaG correlations (1996 History 1607 papers 1 & 2 and coursework with 1996 Business Studies 1351 coursework) the order of magnitude of correlations between SPaG marks falls, to 0.36, 0.31 and 0.43 respectively. Lower correlations here might be valid if SPaG marks are intended to reflect their curricular settings, by giving special weight to technical vocabulary for instance. But if examiners are influenced by subject achievement when awarding SPaG marks, lower correlations could also reflect less than perfect correlation in candidates' performance in different subjects.

- Do correlations between SPaG and subject content tell us anything about the validity of our measurements? Are the patterns of relationships as might be expected from our 'constructs' of the variables concerned or not? The data for the History 1607 examinations in two different years are consistent. That for 1996 is perhaps the more interesting, as it includes data for the coursework element as well as for externally examined papers. Whilst the correlations between SPaG marks in different components seem as high as might reasonably be expected they are lower than the correlations between SPaG and subject content in either the same component or (to a lesser degree) in other components. Correlations between subject content marks in different components are higher still. This kind of pattern is replicated again in the data for 1995 Physical Education 2367. An additional feature here is the availability of data concerning relationships with candidates' grades in GCSE English and Mathematics. The paper 1 and 2 Physical Education examiners' brief SPaG ratings correlate quite well with GCSE English grades (0.58 and 0.54 respectively), thus providing some 'convergent' evidence of concurrent validity. They correlate (only) slightly less well with GCSE mathematics (0.55 and 0.5). Might this be seen as 'discriminant' evidence? Again correlations involving subject content marks are higher still. We should also note that grades in Maths and English are themselves quite highly correlated (0.66). But does any of this tell us much about the underlying relationships between such variables, when SPaG marks are inherently 'weak' variables, based on a handful of marks, whilst the marks available for subject content suggest that they have enjoyed twenty times the measurement effort? Marks for questions may provide fairer comparisons.

Question level comparisons

An alternative basis for comparison is to select those individual questions having similar maximum marks to SPaG assessments and to examine relationships between these variables and SPaG marks, thus providing an approximately level playing field. Suitable question-level data were available for two examinations only and the relevant analyses are displayed in tables 2.3 (relating to 1995 Physical Education 2367) and 2.4 (relating to 1994 History 1607). For Physical Education the sub-questions having a maximum mark of 4 are included alongside SPaG marks (maximum 3) for each paper. For History those sub-questions having a maximum mark of 3 are included alongside SPaG marks for both paper 1 (maximum 5) and paper 2 (maximum 3). The tables provide means and standard deviations for all variables included and the pairwise correlation matrices.

Table 2.3 1995 Physical Education (2367): Correlations between selected sub-questions and SPaG marks

	P1Q1C	P1Q2C	P1Q4C	P1Q5C	P1Q6C	P2Q1C	P2Q2C	P1SPG	P2SPG	Mean	S.D.	Max Mk
P1Q1C	1.00	.49	.24	.37	.37	.24	.41	.45	.42	1.57	1.10	4
	(323)	(133)	(38)	(186)	(158)	(224)	(99)	(322)	(318)			
P1Q2C		1.00	.45	.22	.35	.28	.30	.43	.38	2.10	1.47	4
		(319)	(43)	(191)	(187)	(211)	(129)	(318)	(311)			
P1Q4C			1.00	.29	.40	.33	.33	.43	.26	1.35	1.13	4
			(187)	(69)	(95)	(123)	(62)	(187)	(180)			
P1Q5C				1.00	.29	.25	.35	.35	.26	1.46	1.41	4
				(452)	(269)	(315)	(142)	(452)	(439)			
P1Q6C					1.00	.21	.32	.48	.37	.95	1.26	4
					(421)	(290)	(150)	(421)	(408)			
P2Q1C						1.00	.29	.30	.33	2.08	1.32	4
						(489)	(66)	(482)	(489)			
P2Q2C							1.00	.43	.50	2.18	1.43	4
							(234)	(229)	(234)			
P1SPG								1.00	.47	1.15	.72	3
								(711)	(691)			
P2SPG									1.00	1.18	.73	3
									(703)			

Table 2.4 1994 History 1607: Correlation between selected sub-questions and SPaG marks

	P1Q1A2	P1Q2A2	P1Q3A2	P1Q4A2	P1Q6A2	P1SPG	P2SPG	Mean	S.D	Max Mk
P1Q1A2	1.00	.35	.59	.36	.39	.48	.25	1.43	1.15	3
	(337)	(223)	(153)	(173)	(74)	(337)	(337)			
P1Q2A2		1.00	.33	.36	.26	.44	.28	1.17	1.20	3
		(239)	(98)	(104)	(36)	(239)	(239)			
P1Q3A2			1.00	.39	.26	.52	.29	1.42	1.27	3
			(165)	(54)	(16)	(165)	(165)			
P1Q4A2				1.00	.64	.47	.14	1.99	1.24	3
				(186)	(20)	(186)	(186)			
P1Q6A2					1.00	.46	.18	1.16	1.07	3
					(81)	(81)	(81)			
P1SPG						1.00	.50	2.59	.94	5
						(359)	(359)			
P2SPG							1.00	2.05	.65	3
							(359)			

Even though the variance of both SPaG assessments was less that in the subject content assessments, the Physical Education question-level data correlation matrix shows that the 0.47 correlation between the paper 1 and paper 2 SPaG marks (marked by different examiners) was higher than all 7 correlations between these two SPaG marks and sub-questions on the 'other' papers (marked by different examiners) and 6 out of the 7 correlations between SPaG marks and sub-questions on the 'same' papers (marked by the same examiners). This provides strong evidence of both convergent and discriminant validity. The inter-SPaG correlation coefficient was also higher than all 21 sub-question

inter-correlations, suggesting that, mark for mark, SPaG assessments were perhaps more reliable than assessments of subject content.

In the History matrix (where question level data are only available for paper 1) the variance of the two SPaG assessments was again less than that of all the sub-questions selected. Again however the correlation between the SPaG marks (0.5) was higher than all 5 correlations between paper 1 SPaG and paper 1 questions (marked by the same examiner) and all bar one of the 5 correlations between paper 2 SPaG and paper 1questions (marked by different examiners). So here too we have strong convergent and discriminant evidence for the validity of SPaG assessments. The inter-SPaG correlation here was higher than 8 of the ten correlations between sub-questions, again indicating that, mark for mark, the SPaG assessments were as or more reliable than those of subject content.

However in the matrices for both History and Physical Education, correlations between sub-question marks and SPaG marks in the same paper, hence marked by the same examiner, were however always higher than those for the same sub-questions and SPaG marks from the other paper awarded by a different examiner. This might indicate that 'halo effects' were operating, whereby examiners' marks for SPaG were not fully independent of those for subject content and were influenced by their general perception of the candidates' abilities.

What does it mean to say that SPaG assessments were as reliable as sub-questions carrying similar maximum marks? Table 2.5 illustrates the 'typical' levels of agreement in operation, by presenting cross tabulations of (a) the SPaG marks awarded (by different examiners) for papers 1 and 2 in Physical Education 2367 and (b) the marks awarded (again by different examiners) to candidates who chose to attempt both question 2c in paper 1 and question 2(c) in paper 2 (the correlation between these being close to the median of the values observed between questions in the two papers in this examination).

Table 2.5 Association between (a) SPaG marks in papers 1 and 2 and (b) paper 1Q2(c) and paper 2 Q2(c) marks in 1995 GCSE Physical Education 2367

(a)

		P2 SPaG mark (n (%))				
		0	1	2	3	total
P1 SPaG mark	0	51 (7.4%)	62 (9%)	8 (1.2%)		121 (17.5%)
	1	58 (8.4%)	212 (30.7%)	84 (12.2%)	3 (0.4%)	357 (51.7%)
	2	6 (0.9%)	77 (11.1%)	99 (14.3%)	14 (2%)	196 (28.4%)
	3		6 (0.9%)	8 (1.2%)	3 (0.4%)	17 (2.5%)
total		115 (16.6%)	357 (51.7%)	199(28.8%)	20 (2.9%)	691 (100%)

(b)

		P2 Q2(c) (n (%))					
		0	1	2	3	4	total
P1 Q2(c) mark	0	1 (0.8%)	3 (2.3)	2 (1.6%)	3 (2.3%)	6 (4.7%)	15 (11.6%)
	1	7 (5.4%)	11 (8.5%)	2 (1.6%)	3 (2.3%)	4 (3.1%)	27 (20.9%)
	2	6 (4.7%)	3 (2.3%)	4 (3.1%)	9 (7%)	1 (0.8%)	23 (17.8%)
	3		3 (2.3%)	3 (2.3%)	1 (0.8%)	9 (7%)	16 (12.4%)
	4	3 (2.3%)	4 (3.1%)	8 (6.2%)	9 (7%)	24 (18.6%)	48 (37.2%)
total		17 (13.2%)	24 (18.6%)	19 (14.7%)	25 (19.4%)	44 (34.1%)	129 (100%)

Inspection of the cross-tabulations for the two SPaG ratings and the marks on the two questions reveals that agreement (i.e. broadly, the extent to which candidates are concentrated towards the main diagonals) is much weaker in the latter. This indicates that the examiners for the two questions, considered here as examples of repeated measurements of subject content, disagree more often than examiners did when awarding SPaG marks.

Achieved weight

In practice the various components making up an examination do not always realise their intended weights. The variances and reliabilities of each element and their inter-correlations all play a part in the weight achieved in practice and various models have been suggested to estimate achieved weight. Fowles (1974) advocates the use of component with aggregate covariance, thus taking into account relative component variances and correlations with total marks. This model enjoys the property of additivity and is easily interpreted. Achieved weights sum to 1.0 and component achieved weightings can be interpreted as the proportion each contributes to the total. The 'achieved weight' (or contribution to the final ordering of candidates) of SPaG marks within each component compared to their 'intended weight' (as indicated by the ratio of SPaG to subject content marks) is shown in table 2.6.

Table 2.6 - Intended vs. achieved weight of SPaG marks

Year and syllabus	component	grade range	n	**SPaG intended weight**	**SPaG achieved weight**
1994 History 1607	paper 1	G-A*	359	5.06%	4.00%
	paper 2	G-A*	359	4.76%	4.56%
1995 English Literature 1512	paper 1 (B)	G-D	398	5.06%	4.55%
	paper 2 (A)	F-A*	409	5.06%	4.22%
	paper 2 (B)	F-A*	764	5.06%	4.43%
1995 English Literature 1513	paper 1	G-D	686	5.33%	5.08%
	paper 2	F-A*	732	5.33%	4.46%
1995 Geography 1576	paper 1	G-C	338	4.76%	3.40%
	paper 2	E-A*	332	4.76%	2.68%
1995 Physical Education 2367	paper 1	G-A*	692	4.76%	4.27%
	paper 2	G-A*	692	4.76%	4.20%
1996 Business Studies 1351	coursework	G-A*	1761	5.06%	4.03%
1996 Geography 1588	paper 1	D-G	637	4.76%	2.22%
	paper 2	E-B	597	4.76%	3.28%
	paper 3	D-A*	640	4.76%	2.04%
	paper 4	G-C	618	4.76%	3.36%
	paper 5	E-A*	586	4.76%	3.46%
1996 History 1607	paper 1	G-A*	1634	5.06%	3.91%
	paper 2	G-A*	1634	4.76%	4.61%
	coursework	G-A*	1314	4.76%	4.80%
1996 Religious Studies 1730	paper 1	G-A*	605	4.76%	4.06%
	paper 2	G-A*	614	4.76%	3.33%
	paper 3	G-A*	614	4.76%	4.87%
	paper 4	G-A*	513	4.76%	4.03%
	paper 5	G-A*	608	4.76%	4.58%
	paper 6	G-A*	619	4.76%	3.84%
	paper 7	G-A*	588	4.76%	3.72%
1996 Drama 2325	paper 1	G-A*	589	5.17%	5.31%
	paper 2	G-A*	583	5.17%	6.06%

It should be noted that the nature of SCAA's SPaG regulations means that the intended weight of SPaG assessments is rarely exactly 5%, although it must fall between 4.5% and 5.49%. In only 4 of the 29 cases here do SPaG assessments realise or exceed their full intended weighting. The exceptions are 1996 Drama 2325 papers 1 and 2, 1996 History 1607 coursework and 1996 Religious Studies 1730 paper 3. These are all cases where the SPaG marks are particularly widely spread over the range available, especially by comparison with marks for subject content in the same papers.

In most instances the under-weighting is insufficient to give rise to any real concern. The general tendency for SPaG marks to be relatively well spread (which appears a highly desirable attribute in this light) seems to be sufficient to compensate for their measuring traits rather different from the larger subject content part of their 'host' examinations, which would be likely to contribute to under-weighting.

In a few cases under-weighting is more serious, notably 1995 Geography 1576 (especially paper 2, where SPaG marks achieve only about half their intended weight) and 1996 Geography 1588 (especially papers 1 and 3 where less than half the intended weight is achieved). In both of these examinations the SPaG marks exhibit little spread, probably at least in part because these two syllabuses involved greater differentiation than others included in this review; each paper catering for a relatively narrow range of candidate ability. If examiners are seeing only a comparatively narrow range of candidates they may find it especially challenging to assess language skills, or they may be influenced by the candidates' subject knowledge, which will have contributed strongly to deciding which tier they should enter.

The stability of the statistical characteristics of SPaG assessments

These data include one instance only where we can compare the statistical characteristics of SPaG assessments in the same syllabus in different years: the 1994 and 1996 History 1607 examinations. How stable do these appear? The syllabus and structure of the History 1607 examination papers was unchanged and the size and nature of the candidate entry stayed relatively consistent between 1994 and 1996 (Joint Council 1994 & 1996). The SPaG mark distributions across years were very similar. The mean mark for SPaG in paper 1 was the same across years whilst that for paper 2 decreased slightly between 1994 and 1996. Interestingly this pattern is mirrored by the comparisons in mean marks for subject content; thus supporting theories that examiners are influenced by subject performance in assessing SPaG. The correlation between the SPaG marks was however lower in 1996 (0.41) than 1994 (0.5) but correlations between SPaG marks and subject content were similar, as were achieved weights.

The statistical characteristics of SPaG marks in this syllabus were thus very similar across years.

3 A study of the concurrent validity of SPaG marks

Methodology
SPaG marks are intended to reflect spelling, punctuation and grammar. But candidates may vary with respect to each of these. Might one or two of these three elements dominate? Or are SPaG marks an evenly balanced measure of spelling, punctuation and grammar? Conversely are we sure SpaG measures any of these traits? Evidence (of a relationship between SPaG and the score for subject content) reported above suggests examiners might be influenced by candidates' ability in the subject concerned when awarding SPaG marks. What other 'invalid' factors might influence SpaG marks? For example, could tier of entry, or handwriting, or length of answer be an influence? To consider such issues, independent measures of candidates' achievement in the traits the SPaG mark is intended to measure are required. These might be expected to correlate positively with the marks awarded. As well as such evidence (of convergent validity) it is also desirable to explore the possibility that other, less valid, factors might influence the marks awarded and to seek evidence (of discriminant validity) that this is not the case.

To this end, a random sample of 100 1996 Geography 1588 candidates entering for paper 4 and a further sample of 100 entering for paper 5 were identified. The relevant scripts were obtained for 195 of these; five being unobtainable for administrative reasons. The scrutiny of text to identify prose errors is time consuming and our resources were insufficient for us to consider the whole of each candidate's script. Following similar methodology to that of Massey and Elliott (1996), writing samples were obtained from these scripts. These consisted of the first three sentences of question 7 on paper 4 for lower tier candidates and the first three sentences of question 5 on paper 5 for higher tier candidates. These questions were chosen because they were very similar, with only slight differences between their wording. The writing samples for each candidate were cut out of the scripts and pasted on plain white sheets of paper. They provided a basis on which to estimate each candidate's achievements in spelling, punctuation, grammar, handwriting quality and the length of their answer to this question. In addition, the marks awarded for subject content, the candidates' SPaG marks in the paper and their tiers of entry were noted.

- *Spelling* achievement was estimated by (manually) counting spelling errors (excluding repeat errors) in the sample of writing and dividing by the number of words in the sample, thus obtaining the proportion of words in error. These were then subtracted from 1, so that better spelling was indicated by a higher score (to avoid negative relationships).

- *Punctuation* achievement was estimated by (manually) counting punctuation errors in the sample and dividing by the number of words. The punctuation error count included the incorrect use (or non-use) of full stops, commas, semi-colons, colons, apostrophes and case. Again error 'proportions' were subtracted from 1, so that higher scores represent better performance.

- *Grammar* achievement in grammar was estimated by (manually) counting correctly constructed sentences in the writing sample.

- *Handwriting* quality was estimated by calculating an average rating[1] from independent judgements made by ten people, who had provided impression ratings of the

[1] The median correlation between raters (all UCLES Research & Evaluation Division staff) was 0.52.

handwriting quality of each candidate by sorting them into five piles, ranging from best to worst.

- *Length* The number of words taken to answer the question was estimated by counting the number of words in the sample, dividing by the number lines in the sample and multiplying by the number of lines in the candidate's whole answer for the question.

Were SPaG marks measuring spelling, punctuation and grammar?
The data for candidates entering the two tiers in the examination were considered separately, because marks for subject content (for papers 4 and 5 respectively) were on different scales. Initial analyses calculated means, standard deviations and product-moment correlations (tables 3.1a and b). Subsequent analyses fitted multiple regression models, to review the ways in which variables combine to 'predict' SPaG marks (tables 3.2a and b), and attempted to provide more parsimonious summaries of the correlational structure via factor analyses (tables 3.3a and b).

Table 3.1a - Means, standard deviations and correlations between SPaG, spelling, punctuation, grammar, handwriting, subject content and length of answer for Geography 1588 P4 (Lower Tier)

	Mean	SD	Spelling	Punctn	Grammar	Handwrtg	Subj cont	Length
Spelling	0.96	0.03	-					
Punctuation	0.97	0.03	.17	-				
Grammar	1.65	1.04	.47**	.25*	-			
Handwriting	2.91	0.72	.33**	.31**	.17	-		
Subject content	30.89	10.58	.31**	.24*	.25*	.01	-	
Length	251.66	105.87	.21*	.08	.06	.19	.57**	-
SPaG	1.29	0.80	.46**	.16	.17	.07	.46**	.30**

* sig < 0.05 ** sig < 0.01 (2 tailed) (n = 97)

Table 3.1b - Means, standard deviations and correlations between SPaG, spelling, punctuation, grammar, handwriting, subject content and length of answer for Geography 1588 P5 (Higher Tier)

	Mean	SD	Spelling	Punctn	Grammar	Handwrtg	Subj cont	Length
Spelling	0.98	0.02	-					
Punctuation	0.98	0.03	.36**	-				
Grammar	2.04	1.03	.45**	.21*	-			
Handwriting	2.81	0.83	.29**	.16	.21*	-		
Subject content	31.59	9.81	.30**	.02	.17	.03	-	
Length	312.29	136.63	.34**	.35**	.26**	.17	.44**	-
SPaG	2.33	0.77	.33**	.04	.14	.18	.53**	.29**

* sig < 0.05 ** sig < 0.01 (2 tailed) (n = 98)

How do SPaG marks relate to the 'independent' measures of spelling, punctuation and grammar derived from samples of students' writing? In both tiers, the correlation of SPaG marks with spelling is significant, whilst correlations with punctuation and grammar are not. SPaG marks are correlated with both the marks for subject content and the length of candidates' answers in both tiers; as are the estimates for spelling. Correlations between the estimates for spelling, punctuation and grammar are significant (with the exception of punctuation/spelling in the Lower Tier). The quality of handwriting is correlated with spelling and punctuation estimates in the Lower Tier and with spelling and grammar in the Higher Tier. The length of candidates' answers was related to subject content and SPaG

marks as well as spelling in the Lower Tier and to subject content and SPaG marks, spelling, punctuation and grammar in the Higher Tier.

What can we make of this? Two further statistical approaches summarising these data were used. The first involved the use of a multiple regression model to predict SPaG marks from the spelling punctuation and grammar counts, the handwriting judgements, answer length and marks for subject content. Tables 3.2a and 3.2b provide the results of these in terms of the improvement in SPaG marks associated, on average, with a rise of one standard deviation in 'scores' on each predictor variable. It is evident that in the Lower Tier spelling (but not punctuation or grammar) is associated with increasing SPaG marks, as are marks for subject content. In the Higher Tier the same variables are again the strongest predictors but here it is subject content which predominates. In both tiers the contributions of other variables are trivial and the regression model accounts for 29% of overall variance in SPaG marks. If this appears low we should remember that the criterion (SPaG marks) is a single rating on a 0 - 3 scale, inevitably including a substantial error component which will restrict the predictable percentage of total variance. For instance if the reliability of SPaG marks were about 0.55 (higher than any of the correlations between SPaG marks in different components we have observed) then the maximum percentage of criterion variance we could hope to predict would be about 30%. So whilst the possibility remains that other variables, not included here, are important in 'explaining' SPaG performance, or that the various measures used here are themselves lacking in reliability and/or validity, it seems likely that candidates achievements in the host subject (here geography) and their ability to spell correctly, which are themselves likely to be positively correlated, were the major determinants of the SPaG marks awarded.

Table 3.2a - Multiple regression results for Geography 1588 P4 (Lower Tier) SPaG explained by spelling, punctuation, grammar, handwriting, subject content mark and length of answer

An increase in one standard deviation of	would lead to an increase of ..... in SPaG (out of 3)
Spelling	0.34
Punctuation	0.05
Grammar	-0.08
Handwriting	-0.07
Subject content	0.25
Length	0.04
Adjusted R square = 0.29	n = 97

Table 3.2b - Multiple regression results for Geography 1588 P5 (Higher Tier) SPaG explained by spelling, punctuation, grammar, handwriting, subject content mark and length of answer

An increase in one standard deviation of	would lead to an increase of ..... in SPaG (out of 3)
Spelling	0.14
Punctuation	-0.05
Grammar	-0.04
Handwriting	0.10
Subject content	0.36
Length	0.03
Adjusted R square = 0.29	n = 98

The second summary of the inter-relationships between these variables used principal components factor analysis, followed by varimax rotation of the factors extracted with eigenvalues greater than 1.0 (two in each case), to provide loadings for each variable on common factors. It is worth noting that even oblique rotation produced very similar patterns. The varimax rotated factor loadings are provided in tables 3.3a and 3.3b, where the more substantial loadings (using the usual >0.3 rule of thumb) are picked out in bold type. Note that factors 1 and 2 in the higher tier factor matrix have been transposed to make comparisons between tiers easier.

Table 3.3a Factor Analysis: varimax rotated factor matrix for Geography 1588 P4 (Lower Tier)

	factor loadings	
	factor 1	factor 2
	subject ability	*language*
Spelling	**0.40**	**0.66**
Punctuation	0.08	**0.62**
Grammar	0.16	**0.67**
Handwriting	-0.05	**0.70**
Subject Content	**0.86**	0.10
Length	**0.77**	0.01
SPaG	**0.70**	0.22

Table 3.3b Factor Analysis: varimax rotated factor matrix for Geography 1588 P5 (Higher Tier)

	factor loadings	
	factor 2	factor 1
	subject ability	*language*
Spelling	**0.36**	**0.71**
Punctuation	-0.08	**0.73**
Grammar	0.16	**0.64**
Handwriting	0.04	**0.55**
Subject Content	**0.89**	0.03
Length	**0.53**	**0.46**
SPaG	**0.81**	0.07

This approach too reinforces the similarities in the patterns of relationships between the two tiers. In both case two factors are extracted which can reasonably be described as *subject ability* and *language* respectively. In the Lower Tier data the subject ability factor is the stronger and whilst it attracts a significant loading from spelling (alone of the independent estimates of language skills) it is dominated by the loadings from subject content marks, length of answers and SPaG marks. SPaG marks fail to load substantially on the language factor, reflecting the strength of their association with candidates' overall achievement in the paper. In the Higher Tier the picture is markedly similar except that loadings for length of answers are split. In both cases handwriting quality is associated with the language skills estimates, perhaps reflecting the fact that these all stem from the same data source, even though the two sets of ratings were produced independently.

- It thus appears likely that marks awarded for SPaG are influenced by spelling much more than by punctuation and grammar. If all three elements are regarded as equally important we should perhaps reflect on the validity of the SPaG ratings. We will return to the issues raised by this later.

- We noted previously that the association between SPaG and subject content marks suggests halo effects' may be in operation, whereby examiners are influenced by their

overall view of the candidates. This may or may not be a valid interpretation. SPaG assessments are by design set in a subject context and may emphasise subject specific language, making some association with content marks reasonable, especially as we can expect that candidates who write well will be able to construct effective answers and for there to be a general association between language skills and learning in geography. Such associations might thus be both natural and unavoidable. But the relatively weak association between SPaG and the language factor here is not encouraging and may lend weight to halo effect interpretations.

- It is perhaps reassuring that neither subject content nor SPaG marks were much influenced by handwriting. The correlations with length of answer observed are understandable; with those providing most information winning higher marks.

The equivalence of SPaG assessments in different tiers/options

SCAA's criteria for SPaG assessments in GCSE are universal, applying to all subjects and all tiers of entry. Does this mean that examiners (marking examinations set by different examining groups, in various syllabuses – each perhaps involving alternative tiers of entry and/or optional papers) should all be applying the same 'standards' whilst awarding SPaG marks?

We will focus on one aspect of this question here, the equivalence of SPaG marks across tiers in the same examination. Table 3.4 reproduces the mean (proportion of maximum) SPaG marks awarded to candidates in the various papers in all the tiered examinations for which we have data. It is clear that SPaG marks achieved in higher tiers are, on average, markedly better than those on lower tiers. This suggests that examiners in each tier are certainly not simply spreading their own candidates across the full range of SPaG marks and suggests that to some extent at least they are detecting the 'quality' of higher/lower tier candidates and rewarding them accordingly. But are they doing so accurately, in the sense that they can make the appropriate allowances and award equivalent SPaG marks for equivalent performance? Any attempt to answer this question requires more information, in the form of some common yardstick(s) against which candidates in different tiers can be compared.

Again the supplementary data concerning 1996 Geography 1588 candidates ability in spelling punctuation and grammar gathered from the writing samples, as described above, helps us to address the issue more directly. If similar standards are being applied by examiners in papers 4 and 5 (the Lower and Higher Tiers respectively) it would seem reasonable to expect that candidates with similar 'scores' in spelling, punctuation and grammar derived from their writing samples should, on average, be awarded similar SPaG marks. What happens in practice?

Table 3.4 - Mean SPaG proportions in tiered syllabuses

Year and syllabus	component	grade range	n	SPaG mean proportion
1995 English Literature 1512	paper 1 (B)	G-D	398	.48
	paper 2 (A)	F-A*	409	.74
	paper 2 (B)	F-A*	764	.70
1995 English Literature 1513	paper 1	G-D	686	.49
	paper 2	F-A*	732	.73
1995 Geography 1576	paper 1	G-C	338	.45
	paper 2	E-A*	332	.81
1996 Geography 1588	paper 1	D-G	637	.19
	paper 2	E-B	597	.43
	paper 3	D-A*	640	.87
	paper 4	G-C	618	.38
	paper 5	E-A*	586	.82

Fitting a multiple regression model in which SPaG is explained by spelling, punctuation, grammar and tier shows us how far tier of entry affects SPaG marks after controlling for variations in spelling, punctuation and grammar. Given that previous analyses revealed the likelihood that subject content marks are important in determining the SPaG marks awarded, a second regression model was also fitted which includes this additional variable. Table 3.5 summarises these analyses for both models, revealing that Higher Tier candidates were, on average, likely to gain 0.89 more SPaG marks than Lower Tier candidates who had 'equivalent' levels of achievement in spelling, punctuation and grammar. When subject content marks are added to the model, so improving the fit (with R^2 rising from 0.41 to 0.51), the estimated effect size increased marginally, so that higher tier candidates appeared to gain an average mark 0.92 greater than their equivalents in the Lower Tier. The evidence suggests that the standard of spelling, punctuation and grammar required to gain a given SPaG mark is not consistent across the two tiers, thus challenging the validity of any claim that SPaG assessments are made on a single scale.

Table 3.5 - The tier coefficient in multiple regression analysis (1996 Geography 1588 Papers 4 & 5)

model	tier coefficient (d)	R square
SPaG = (a)*spelling + (b)*punctuation + (c)*grammar + (d)*tier + constant	0.89	0.41
SPaG = (a)*spelling + (b)*punctuation + (c)*grammar + (d)*tier + (e)*paper score + constant	0.92	0.51

n = 195

Does this matter? Are candidates disadvantaged if, as it appears, different marking standards are used in SPaG assessments in different tiers (and optional papers, as if different standards are used here this may also be so in other circumstances)?

Where the structure of the examination is such that candidates taking one tier/option are graded quite separately from those taking other options (e.g. where candidates take either papers 1 and 2 or 3 and 4), variations in SPaG marking are clearly no more likely to affect the grades awarded to candidates from different options who have equivalent language skills than they are to affect the grades awarded for candidates taking different syllabuses or subjects or examinations set by different boards: so no-one is disadvantaged.

But what of syllabuses where the candidates may choose between alternative optional papers (e.g. where candidates all take paper 1, plus either paper 2 or paper 3)? Is this really any different? Even here it is not obvious how candidates in options subjected to more stringent marking standards will be disadvantaged. Problems would only arise if examiners failed to take any variations in SPaG marks awarded into account when judging the minimum mark thresholds for the award of each grade, as SCAA's mandatory code for the conduct of GCSE examinations requires they should. But it is difficult, and perhaps impossible, to say if this requirement can be successfully achieved. Those judging grading standards will not normally have analyses like ours to help them see if and how SPaG marking is awry. Typically only the distributions of candidates' total marks on each paper (including SPaG) are available in the course of operational examining. Examiners must thus try to disentangle the effects of varying SPaG marks amidst the cut and thrust of discussions about the quality of work in individual scripts from small groups of candidates awarded each total mark for the component, which provide the key evidence for standard setting judgements.

So whilst in theory the procedures for grading suggest that there is no reason why variations in SPaG marks between options will result in inequities in grades, there is certainly a risk that variations in SPaG marking standards might prove a complicating factor and thus contribute to the difficulty of establishing equivalence between options, if professional judgements are largely the basis for doing so. We should however recognise that professional judgements are not the sole basis for such decisions and that in many instances statistical indicators are also provided to support the decisions of the examiners involved.

4 Gender and SPaG assessment

Do males and females vary with respect to SPaG marks? If so is it fair? To address this we will need to compare the performance of boys and girls, making allowance for any systematic differences in the abilities of the boys and girls entering any particular examination/ tier etc. Males and females may also differ in their aptitudes for different subjects, as well as for SpaG. Where specialist aptitudes are required any apparent sex differences in SPaG marks (which can be seen as being foisted on to the subject by the regulatory body) may be problematic if they are not in line with the patterns of sex differences commonly found in achievement in the specialism itself.

Previous research into SPaG has provided some evidence that females obtain higher SPaG marks than males (Adams 1993a) which may relate to deep-seated sex differences in language skills. Do similar differences exist in the datasets considered in this study? How should they affect SPaG marks?

The mean (proportion of) SPaG marks for boys and girls in the syllabuses for which we have data are shown in table 4.1, as are their mean proportions of subject content marks. Note here that it has already been shown that SPaG marks are influenced by candidates' general performance on the same paper. As such, subject content marks will have to be taken into account when comparing SPaG marks. This can be achieved by incorporating the SPaG marks, subject content marks and gender into a multiple regression model, as shown in equation 4.1.

Equation 4.1 - Regression equation for gender analysis

SPaG = (a) * subject content mark + (b) * gender + constant + error

where gender = 0 if female and 1 if male

The size of the gender coefficient will show the effect of gender, after taking into account subject knowledge. As the maximum available SPaG marks for different syllabuses and components vary, the gender coefficient can be scaled by the maximum SPaG mark so that we can compare SPaG marks for males and females in terms of proportions of available SPaG marks. These coefficients are also shown in table 4.1 for all syllabuses for which we have suitable data.

In nearly all examination components considered in this study, females have higher average SPaG marks than males and some cases the differences appear quite large (e.g. both PE 2367 and Drama 2325, where the gender difference amounts to 0.13 of total SPaG marks). Paper 1 in 1994 History 1607 is the only exception. Thus at first sight females appear to be out-performing males. This might however arise from an overall difference in the calibre of the boys and girls entering for these syllabuses. The mean (proportions of) marks for subject knowledge also exhibit differences between males and females, though these are not so one-sided, with boys doing better than girls in almost one third of these components. To be sure that a good SPaG performance is due to better SPaG rather than just higher ability, we must control for differences in subject content marks in the analysis and the final column in table 4.1 shows the regression coefficient for gender. A positive coefficient means boys are performing better than girls,

whilst a negative coefficient shows that females are performing better than males on SPaG - after controlling for subject ability.

In 25 out of the 29 cases the regression coefficient is negative, showing females are performing better than boys in SPaG after controlling for subject knowledge marks. In about half of these the differences observed are large enough to be statistically significant. There are no instances of boys out-scoring girls to a statistically significant extent. This confirms the tentative findings on sex differences reported by Adams (1993a).

Table 4.1 - SPaG and gender

Year and syllabus	component	grade range	SPaG male mean proportion	SPaG female mean proportion	Subject knowledge male mean proportion	Subject knowledge female mean proportion	Regression coefficient (boys' SPaG score in relation to girls', as a proportion of max SPaG marks. Negative values show boys are performing worse than girls)
1994 History 1607	paper 1	G-A*	.66	.64	.50	.45	-.022
	paper 2	G-A*	.66	.71	.51	.50	-.049*
1995 English Literature 1512	paper 1 (B)	G-D	.45	.51	.33	.35	-.050**
	paper 2 (A)	F-A*	.72	.75	.63	.68	.006
	paper 2 (B)	F-A*	.69	.72	.57	.61	-.005
1995 English Literature 1513	paper 1	G-D	.47	.52	.30	.33	-.021
	paper 2	F-A*	.71	.74	.51	.53	-.009
1995 Geography 1576	paper 1	G-C	.44	.46	.53	.51	-.031*
	paper 2	E-A*	.79	.83	.56	.61	-.014
1995 Physical Education 2367	paper 1	G-A*	.35	.48	.39	.47	-.060***
	paper 2	G-A*	.36	.47	.50	.51	-.104***
1996 Business Studies 1351	coursework	G-A*	.63	.69	.54	.55	-.047***
1996 Geography 1588	paper 1	D-G	.18	.19	.44	.39	-.036***
	paper 2	E-B	.40	.47	.52	.51	-.084***
	paper 3	D-A*	.85	.89	.54	.52	-.038**
	paper 4	G-C	.37	.41	.50	.51	-.037*
	paper 5	E-A*	.81	.82	.55	.53	.001
1996 History 1607	paper 1	G-A*	.64	.65	.50	.47	-.026***
	paper 2	G-A*	.61	.64	.57	.55	-.046***
	coursework	G-A*	.62	.65	.67	.65	-.050***
1996 Religious Studies 1730	paper 1	G-A*	.53	.64	.47	.51	-.083***
	paper 2	G-A*	.56	.62	.39	.47	-.011
	paper 3	G-A*	.52	.58	.49	.54	-.014
	paper 4	G-A*	.62	.66	.53	.58	.011
	paper 5	G-A*	.43	.53	.43	.52	-.014
	paper 6	G-A*	.69	.79	.50	.59	-.025
	paper 7	G-A*	.54	.61	.46	.51	-.041
1996 Drama 2325	paper 1	G-A*	.51	.59	.44	.50	-.021
	paper 2	G-A*	.45	.58	.44	.55	.008

significance levels : * = 5%, ** = 1%, *** = 0.1%

The data derived from writing samples of 1996 Geography 1588 candidates are also of interest in this regard, as the initial analyses of these data explored the importance of spelling, punctuation and grammar in determining SPaG marks. It would appear worthwhile considering the relative distributions of scores of males and females on these variables. Section 3 above showed that it was spelling, rather than punctuation and grammar which seemed most likely to influence SPaG marks. Are the girls better at spelling?

Mean scores for spelling, punctuation and grammar from the samples of writing are given in table 4.2. They show that whilst males and females have similar levels of performance in punctuation and grammar, females made significantly fewer spelling mistakes. This echoes Massey and Elliott (1996), who found girls less likely to make spelling mistakes than boys, but little difference in grammar and most forms of punctuation. Given the predominant influence of spelling on marking, the girls' superiority in SPaG marks is therefore perhaps to be expected. However, if markers were to pay more attention to the punctuation and sentence construction elements we might expect the 'typical' deficit of boys SPaG marks to be reduced, though not necessarily eliminated.

Table 4.2 - 1996 Geography 1588 writing sample: spelling, punctuation and grammar by gender

		mean	
		male (n=102)	female (n=93)
Spelling	proportion of words mis - spelled	0.037	0.023**
Punctuation	number of punctuation errors divided by number of words	0.026	0.023
Grammar	number of correctly constructed sentences (out of 3)	1.863	1.833

** difference statistically significant beyond the 0.01 level

If these findings are generalisable, the requirement to include SPaG in the schemes of assessment for many GCSE subjects, together with this tendency for SPaG markers to reward to males and females differentially, would appear to be a contributing factor to the general tendency for boys to obtain lower GCSE grades than girls in recent years (Stobart et al, 1992), although it is clearly not the only factor involved.

Is it acceptable for SPaG marking to emphasise a feature of language where there are discernible sex differences or is this a 'bias' which should be removed? Or does this complicating factor weigh against the inclusion of SPaG in GCSE examinations willy nilly?

5 The influence of SPaG on candidates' grades

Work on the initial SPaG assessments in 1992 (Adams 1993a) suggested that the grades of quite large numbers of candidates were affected. But the regulations governing the way in which SPaG assessments must be integrated into standard setting have been made explicit since 1992, so that procedures may have changed. Grade boundary decisions must now be based on a judgement about the overall quality of candidates. Scripts and marks incorporate SPaG assessments and so such judgements must take account of the levels of accuracy in spelling, punctuation and grammar 'typical' of students at each grade. But how often does 'untypical' SPaG performance effect the GCSE grade of a candidate?

Methodology
Now that SPaG marks are an integral part of the process, we have to find a means of estimating their impact. We have attempted this by calculating, via regression, the 'expected' SPaG score for candidates awarded a given subject content mark, and substituting these expected scores for their real SPaG marks when calculating an alternative 'new' syllabus total. Candidates who do not conform to the pattern of SPaG marks typical for students of their level of achievement in subject content will thus have a new total mark which differs from their actual total. By comparing the grades which would have been awarded for these two total marks, we can estimate the size of the effect untypical SPaG performance has on the final GCSE grades.

1996 History (1607) was chosen as the syllabus most suitable for this analysis because we had data for a large sample which included the SPaG marks for two written papers and for coursework (n=1314), all of which were taken by all candidates.

The raw[1] subject content total marks awarded for each question on each paper were obtained from scripts and combined with the SPaG marks awarded, scaling each to reflect the element's intended weight in the syllabus total. The grade thresholds were then used to calculate a grade for each candidate.

The SPaG marks for all components were then aggregated to form an overall SPaG total mark, as were marks for subject content (again using appropriate scaling factors in both cases). The total for subject knowledge was then regressed (equation 5.1) on the total for SPaG and the regression equation was used to predict an expected SPaG total mark for each candidate, given their subject content mark. A revised syllabus total was then calculated by adding the predicted SPaG total to the subject content total for each candidate. These new totals were then graded as before, to give the revised grades which might have been awarded if all candidates had performed as 'predicted' in SPaG, given their marks for subject content.

Equation 5.1 - SPaG Regression equation

$$SPaG = 0.046 * (\text{subject knowledge}) + 1.29$$

[1] Examiner scalings, applied to paper totals in some cases to reflect senior examiners' judgements about the relative severity/lenity of asisstant examiners, were excluded from this study because it was impossible to discern whether or not they had arisen from SPaG or content marks.

The grade distributions for grades based on raw SPaG marks and grades based on predicted SPaG marks are shown in table 5.1. The two distributions are very similar. The largest divergence was at a grade A, where there was a difference of 0.5%.

Table 5.1 Grade distributions for raw mark syllabus grade and the syllabus grade including a predicted SPaG mark

	raw mark grade	revised grade based on predicted SPaG marks
A*	3.7%	3.3%
A	15.6%	16.1%
B	28.3%	28.6%
C	22.3%	22.0%
D	9.1%	9.2%
E	6.8%	6.6%
F	6.2%	6.2%
G	5.6%	5.6%
U	2.4%	2.3%

The grade changes induced by untypical SPaG marks are shown in table 5.2. The grades of the vast majority of candidates are unaffected by the substitution of predicted for real performance in SPaG. No candidates are affected by more than one grade. Interestingly, greater proportions of the candidates awarded higher grades were affected, especially at A*, where SPaG marks make a difference for six out of the forty-nine reaching these heights. Few candidates are affected at the lower grade boundaries, despite the fact that the maximum SPaG mark available is a much more substantial proportion of the threshold mark for these grades, which might lead to the supposition that the impact on grading might be stronger here.

Table 5.2 - Grade changes induced by SPaG

		grade based on predicted SPaG mark								
		A*	A	B	C	D	E	F	G	U
	A*	43	6	0	0	0	0	0	0	0
	A	1	200	4	0	0	0	0	0	0
grade based	B	0	5	364	3	0	0	0	0	0
on raw	C	0	0	8	284	1	0	0	0	0
SpaG mark	D	0	0	0	2	117	0	0	0	0
	E	0	0	0	0	3	86	1	0	0
	F	0	0	0	0	0	1	80	1	0
	G	0	0	0	0	0	0	1	72	0
	U	0	0	0	0	0	0	0	1	30

To summarise the effect size, only 2.9% of candidates would have obtained different grades if they had been awarded SPaG marks typical of candidates of their level of subject knowledge instead of those they actually achieved. In all 1.2% of candidates 'improved their grade' by performing better in SPaG than expected, whilst 1.7% of candidates 'reduced' their grade by performing less well in SPaG than expected.

Is this too many or too few or about right? Clearly it is a matter of judgement. Any answer must balance the desire for the outcome of an examination in history to be based primarily on students' knowledge of the subject against the wish to encourage pupils and teachers

to value language skills in the context of teaching and learning history, the primary consequence sought by the introduction of SPaG marks in GCSE.

If this level of influence is generalisable, the impact of SPaG on grading amounts to one candidate per class being affected by a change of plus or minus one grade. This is markedly lower than the impact observed when SPaG assessments were first introduced; perhaps partly because of the changes in arrangements for SPaG since 1992. If teachers and candidates realise that so few candidates are directly affected, it seems unlikely that many will be driven to worry more about accurate language than they do already.

6 Summary and conclusions

- The limited evidence available suggests it is possible that more candidates are likely to be awarded extremely high or low SPaG marks for coursework (as compared with externally examined components catering for the full ability range), perhaps either because some candidates are able to produce better presented work in coursework than in timed examinations or because of the markers' (the candidates' own teachers) closer knowledge of the candidates.

- Distributions of SPaG marks in tiered syllabuses reflected the abilities of the candidates entering each tier, with higher marks being awarded to higher tier candidates. This would seem appropriate, given that all markers were working to the same criteria.

- In papers without tiering, mean SPaG marks were mainly in the range 0.53-0.68 of maximum. The very high levels of SPaG marks often recorded in the first year of SPaG assessments were not found in this review. Standard deviations were all in the range 0.21 - 0.3 of maximum and SPaG marks tended to be proportionately more widely spread than marks for subject content.

- Standard deviations in tiered question papers were often smaller than those in untiered papers. Again however SPaG marks were usually more widely spread than those for subject content.

- Examiners in some subjects were rather loath to award zero marks for SPaG. Others were seemingly less tender minded. But these variations may not just be mere caprice on the markers' part. If markers in different subjects are using the SCAA criteria for SPaG marking it could be hypothesised that, to some extent at least, such distributional variations might relate to self-selection in candidates' entries for different subjects/ options.

- In the one example where suitable data were available, the statistical characteristics of the SPaG assessments involved appeared relatively stable between years.

- Typically, correlations between SPaG and subject content marks (awarded by the same examiners within the same paper) of about 0.6 are observed in externally examined papers catering for the full GCSE ability range. These values are a little higher than those reported from the early days of SPaG assessments. They might be seen as evidence of halo effects, where examiners are influenced by their general impressions when awarding SPaG marks. But we should be cautious about inferring causation from correlation. Association between subject content and SPaG marks may or may not represent lack of validity as the assessment of SPaG has, as a matter of policy, been set in subject contexts and may legitimately emphasise subject specific features, especially vocabulary. It is also inherently reasonable to expect that the effect of general ability on achievement means that good candidates will, typically, have better language skills than weaker ones.

- Correlations between SPaG and content marks are somewhat higher in the examples for coursework available, perhaps either because coursework offers for diligent candidates greater opportunities to earn marks for presentation, including accurate language, or the teachers' assessments embody halo effects.

- In tiered examination papers the restrictions in ability range result in lower correlations between SPaG and subject content marks, echoing previous work, with values around 0.4 being more typical.

- On limited evidence it appears that correlations between SPaG marks awarded by (different) examiners for different papers in the same syllabus are, mostly, in the region of 0.5: higher than those observed in the early stages of SPaG assessments. This seems reasonable evidence of reliability given the brevity of the assessments involved. This contradicts the conclusions reached by earlier researchers and is discussed further below. Again on limited evidence, it seems that correlations between SPaG marks awarded by examiners in different subjects are somewhat lower, probably reflecting the curricular contexts of the assessment of SPaG.

- Patterns within correlation matrices involving both SPaG marks and subject content marks provide some convergent and discriminant evidence that these two forms of assessments are measuring different, albeit related, traits. However the much greater measurement effort (quite properly) attached to assessments of subject content, compared to SPaG, makes it difficult to appreciate the underlying relationships between them.

- Correlations between selected question-level subject content marks and SPaG marks are revealing and suggest that SPaG assessments are probably more reliable, mark for mark, than most examination questions. They also provide impressive convergent and discriminant evidence that the two types of assessment measure different traits. This new approach contradicts the conclusions reached by previous work on both these counts. However these question-level correlations do also suggest that examiners' marks for SPaG are not fully independent of those awarded for subject content, which accords with previous findings; suggesting that the difference of opinion with previous work may be a matter of degree and interpretation, rather than principle or radically different evidence. Cross-tabulations of marks awarded are provided which illustrate the 'typical' levels of agreement between different examiners for SPaG and other questions.

- SPaG marks' achieved weighting is typically marginally below their intended weighting, suggesting that the fact that they measure something rather different from most questions is slightly over-compensating for their typical proportionately wider spread of marks. In most cases this disparity is not large enough to be a source of concern. The cases where under-weighting appears most serious involved examinations where tiering structures created papers targeted at relatively narrow ranges of ability and it seems possible that examiners find SPaG assessment more problematic in such circumstances. As changes in QCA regulations for GCSE have introduced tiering to a wider range of syllabuses it is possible that this will reduce SPaG's overall impact.

- A concurrent validity study (involving a small sample of Geography 1588 candidates) which collected independent supplementary data concerning candidates' skills in spelling, punctuation and sentence construction, indicated that SPaG marks were more heavily influenced by candidates' spelling than by grammar or punctuation. This lends weight to earlier researchers' pessimism about the capacity of brief SPaG assessments to assess such complex constructs. If all three elements are considered equally important, this finding has substantial implications for the validity of current SPaG assessments.

- Arguably, the concurrent validity study also supports the view that halo effects were in operation, as examiners' SPaG marks were more closely related to the marks they awarded for subject content than to concurrent estimates of language skills.

- The concurrent validity study's data also allowed us to explore the issue of the comparability of SPaG assessments across tiers. In the examination concerned the average SPaG marks awarded to candidates rose in higher tiers. But were the examiners making sufficient, or too much, allowance for the variations in ability between the groups of candidates taking each tier? The supplementary data provided a common yardstick against which to measure such abilities and suggested that Higher Tier candidates gained SPaG marks which were on average 0.92 greater than Lower Tier candidates with equivalent spelling, punctuation and grammar skills, so different marking standards may have been applied in the two tiers. However if SPaG quality is considered as an integral part of the process of setting grade thresholds, as is supposed to be the case, candidates should not be disadvantaged, as has been noted before. Incorporating SPaG in threshold setting judgements is however not straightforward and there is no empirical evidence (here or elsewhere) bearing upon this.

- Female candidates obtained higher average SPaG marks than males in nearly all of the examination components included in this review. In some cases the differences appeared quite large (up to 0.13 of maximum SPaG marks). But if the girls entering an examination were more able than the boys this pattern would be the natural result, so analyses must take ability into account. Comparison of the equivalent marks for subject content revealed that these were not so one-sided, with boys doing better than girls on average (excluding SPaG marks) in almost one third of the examination components investigated. A regression model which controlled for achievement in subject content suggested that in 25 out of the 29 components studied, females did in fact obtain higher SPaG marks than males of equivalent ability in the subject and that the differences were statistically significant in about half of these cases: thus confirming previous tentative research findings regarding gender and SPaG.

- The limited evidence from the concurrent validity study confirmed research elsewhere suggesting that girls are superior to boys in spelling, but not in punctuation and grammar. The strong influence which spelling seems to exercise over the award of SPaG marks would thus seem to discriminate against males, contributing to boys' relatively low levels of achievement in GCSE examinations in recent years.

- Study of the impact of SPaG in a GCSE history examination suggested that only about 3% of candidates marks for SPaG were so untypical of candidates of their general level of ability that their overall grades would have been affected. This is much lower than the numbers of candidates estimated to have been affected in 1992, though the changes in the regulations and procedures since then are likely to have done much to bring this about. It is difficult to say if 3% is too low, too high, or about right, without bringing to bear value judgements concerning the extent to which we might wish to influence the behaviour of teachers and pupils. Such consequences are of course central to the purpose of SPaG and hence to judging the value of these assessments. But as teachers and pupils become aware that so few candidates' grades are affected, it would seem likely that SPaG assessments will at best appear a very small carrot, rather than a stick. If so will they continue to serve any useful purpose?

Discussion
This evaluation is less condemnatory than earlier work. In particular it suggests that SPaG assessments are at least as reliable as examiners' other marks and that they are not simply determined by the examiners' general perceptions of the candidates' abilities, though these might have some influence. However this is not to say that still greater reliability (and validity) might not be achieved.

It looks as though the main determinant of SPaG marks is spelling, rather than punctuation and grammar, which may be particularly unfortunate for boys, who tend to be weaker at this than other aspects of accuracy in writing. If the other elements in SPaG are thought to be equally important, ways must be sought to bring them to the examiners' attention. Examiners' competence, and even willingness, to assess these skills is likely to be an issue and better briefing materials would seem essential.

SPaG assessments' typical failure to achieve their 5% expected weighting must contribute to their low impact on grades awarded. This low impact is probably not yet widely realised but over time it is likely that teachers will become less concerned about SPaG than they are at present (and no evidence has yet been produced to show that they have ever been sufficiently concerned to change their classroom practice). This might suggest that other means of raising the profile of accurate writing might be more effective, such as reporting a subsidiary grade for accuracy within English Language GCSE. But some might take the phlegmatic view that if SPaG assessments have little impact they can do little harm and there is little point in tinkering with them, given that they were only brought into being to flag the importance attached to these skills, rather than to change candidates' grades in, say, geography. Let us be grateful for serendipity and leave well alone, they might say.

Attractive as this laissez-faire approach may be, we cannot subscribe to amateurism in assessments which matter so much to so many young people. If a positive political decision to abandon SPaG assessments in favour of some other course is not forthcoming, they should be made to work as effectively as possible. But any improvements are likely to place greater emphasis on assessing SPaG than the present methods, which would certainly take up more of examiners' time (and thus costs). There are certainly likely to be strict limits to the resources we will wish to devote to this purpose.

At this point we find it impossible not to speculate briefly, though probably prematurely, about ways in which SPaG assessments within subject examinations might be improved (on the assumption that national policy continues to dictate their continuation in much their present form).

- The dominance of spelling could be redressed by requiring it to be assessed explicitly, in the form of a separate rating for spelling. This would leave the other two elements in need of assessment in their turn. Might emphasising grammar and punctuation within 'quality of expression' or 'structure and style', in marking instructions encourage non-English specialist examiners to address these other elements impressionistically, in a single rating, rather than to continue to ignore them for fear their judgements might be wrong? The two ratings would then need to be combined and this approach would add to the labour in SPaG assessment, but it would probably be feasible. This seems likely to be compatible with the revisions to GCSE criteria envisaged for 2003 (QCA, 2000) which propose the (optional) inclusion of assessments of 'suitable structure and style in writing'.

- If any further enhancement of reliability is needed (and in our view it is not), might a limited number of repeated estimates, each focused on selected 'rich' sources of evidence within the scripts (i.e. selected questions), be the best way forward?

- Might the production of a generic approach to SPaG assessments which can be used in a wide range of subjects, backed up by high quality support materials for examiners, be the way of securing better assessment, instead of leaving it to the examiners in each syllabus to devise their own salvation? This would encourage marks to reflect candidates' self-selection into different subjects and/or options and should help examiners involved in setting standards, though it might also restrict the range and hence the achieved weightings of SPaG marks for tiered papers. Such a development could be undertaken on a national scale, sponsored by QCA, or by each examining body.

- Examiners' interest in assessments of this nature might be enhanced by increasing the emphasis on the 'subject context' of writing. Again this is not incompatible with the remaining requirement proposed for 2003, to 'present relevant information in a form that suits its purpose'. But such general criteria will need some amplification before subject specialists can make effective assessments of this sort.

Further research and development is needed. Replication would be of value and we recognise that this study has considered many aspects of these assessments, but it has not evaluated the effects of different examiners, which might be of great interest.

References

ADAMS, R.M. (1993a) The assessment of spelling, punctuation and grammar in GCSE examinations in 1992: A report by the Inter-Group Research Committee (IGRC).

ADAMS, R.M. (1993b*) A statistical review of SPG marks in GCSE English, English Literature and History A 1993* Cardiff: Welsh Joint Education Committee.

FOWLES, D.E. (1974) CSE: two research studies, *Schools Council Bulletin* 28, London: Evans/Methuen.

JOINT COUNCIL FOR THE GENERAL CERTIFICATE OF SECONDARY EDUCATION (1994) *Inter-group statistics - summer 1994,* Guildford: Southern Examining Group.

JOINT COUNCIL FOR THE GENERAL CERTIFICATE OF SECONDARY EDUCATION (1996) *Inter-group statistics - summer 1996*, Guildford: Southern Examining Group.

MASSEY, A.J. (1983) The effects of handwriting and other incidental variables on GCE A level marks in English Literature*. Educational Review*, 35,1,45-50.

MASSEY, A.J. and ELLIOTT, G.L. (1996) Aspects of writing in 16+ English examinations between 1980 and 1994, *Occasional Research Paper 1*, Cambridge: University of Cambridge Local Examinations Syndicate.

MEG (1992*) IGRC Study on Spelling, Punctuation and Grammar,* Cambridge: Midland Examining Group.

OWEN, R. (1992) Why SPG is a travesty of justice, *Times Educational Supplement*, 7.8.92.

QUALIFICATIONS AND CURRICULUM AUTHORITY (2000) Common Criteria for GCSE, London: Qualifications and Curriculum Authority.

SCHOOLS CURRICULUM AND ASSESSMENT AUTHORITY (1995) *GCSE Mandatory Code of Practice*, London: Schools Curriculum and Assessment Authority.

SCHOOLS CURRICULUM AND ASSESSMENT AUTHORITY (1997) *GCE A & AS Code of Practice*, London: Schools Curriculum and Assessment Authority.

SCHOOLS CURRICULUM AND ASSESSMENT AUTHORITY (1996) *Key Stage 2 Tests 1996 English Mark Schemes*, London: Schools Curriculum and Assessment Authority.

STOBART,G., ELWOOD,J. and QUINLAN,M. (1992) Gender bias in examinations: how equal are the opportunities?*, British Educational Research Journal*, 18, 3, 261-276.