NUMBERING NESTED QUESTIONS

Sofia Parkinson and Neil Stringer

SUMMARY

In line with the aim of simplifying and standardising the format of AQA's question papers, it has been proposed that all question papers should be numbered using a simple sequential system (i.e. 1, 2, 3… etc.). The study reported here investigated whether using sequential question numbering where an alphanumeric system (1(a)i, 1(a)ii,… etc.) is currently used might have any impact on candidates. 485 Year 11 GCSE Geography candidates participated in the study. Approximately half of the candidates sat a mock examination of the original, unmodified question paper, which used alphanumeric question numbering throughout the three sections; the remaining candidates sat a modified paper in which all but those questions in the first section were renumbered sequentially. Candidates' performances in the first, common section were used as a covariate when comparing the performance of the two groups of candidates on the differently-numbered sections. The analysis showed that the type of question numbering system used by candidates had no significant effect on their performance overall. Likewise, the facility indices of all part-questions were very similar for both groups of candidates and an analysis of candidates' rubric infringements did not find any evidence of systematic error caused by the sequential numbering system. Lastly, candidates in the modified paper condition were asked to comment on their experience of using the two question numbering systems in the mock examination. They identified strengths and weaknesses of both the alphanumeric system and the sequential system but expressed no strong opinion as to which system they preferred. It was concluded that the decision whether or not to discontinue using the alphanumeric question numbering system can probably be made without any concern that candidates will be disadvantaged either way.

INTRODUCTION

The move towards Computer Marking from Image+ (CMI+)[1] for long form answers, such as essays, has raised an issue concerning the way that question papers are numbered. One of the challenges for introducing the CMI+ system for components with long form answers is the segmentation of answer booklets. The system needs to detect not only where an answer to a question starts and finishes but also what question is being answered. Some of AQA's question-numbering conventions are not compatible with the capabilities of the machine readers used for CMI+, at least not to the degree of accuracy that makes automation worthwhile. For example, the use of letters (a, b, c) and Roman numerals (i, ii, iii) to denote part-questions, such as 1(a)i, cannot be accommodated by the system. It has, therefore, been decided that a purely numerical system will be adopted for these papers, such that any item can be identified using a two digit code (01 – 99). Question papers currently in the format of combined question paper/answer booklets, and all question papers that can be put into a combined question paper/answer booklet format for June 2010, may continue to use the nested numbering system, e.g. 1(a)i.

[1] An online system where examiners mark scanned images.

Nonetheless, in the long run, maintaining two parallel question paper numbering conventions where one will suffice is at odds with the goal of promoting a consistent ‗house style" with which candidates and teachers are familiar across specifications and subjects. The current study investigates whether there is any measurable benefit in maintaining the nested numbering system.

GCSE Geography and GCSE Mathematics are two subjects that make extensive use of the nested numbering system because they both tend to use stimuli and scenarios to generate a series of thematically related questions. In some cases, the answer to one part-question may be required to produce the answer to a subsequent part-question. Several senior examiners for GCSE Mathematics expressed concern at the prospect of losing the alphanumeric system although, at that stage, the proposed replacement was a numerical equivalent to the alphanumeric system, where Arabic numerals would replace the letters and Roman numerals, e.g. 1(a)i would become 1-1-1 and 2(c)iv would become 2-3-4. Although (perhaps mercifully) this system was not introduced, the simple sequential numbering system might make the relatedness of these part-questions less obvious to candidates than the alphanumerical system (and even its Arabic equivalent). Presumably, it was this clarity that the examiners were concerned about losing. Unfortunately, owing to specification development commitments, it was not possible to recruit to this study the senior examiners from GCSE Mathematics who had vocally opposed the change; however, a Principal Examiner and his Assistant Principal Examiner were recruited from the GCSE Geography A legacy specification to collaborate on this study.

In terms of how the sequential numbering system might affect candidates in comparison with the alphanumeric system, there are three objective measures that this study focuses on: the time candidates take to complete the questions, their question scores, and the incidence of rubric infringements. Candidates‘ subjective impressions of using the sequential numbering system were also collected, to complement the objective measures. For example, candidates may like the new numbering system but, because they are unfamiliar with it, appear to perform worse when using it; alternatively, the objective measures might show that the numbering system had no affect on performance, but candidates may express strong dissatisfaction with it.

So that the worst case scenario was tested, only the numbering system was altered on the papers used in the study. In practice, it seems likely that alterations to the format of the question papers could, in many cases, be used to compensate for the change in the numbering system and thus offset any apparent disadvantages introduced by it.

METHOD

Participants

The centres with the highest entries for GCSE Geography A (3031) in the academic year 2009/2010 were contacted and asked whether they were planning to hold mock examinations of paper 3031/2 for their year 11 students in the winter. Of those who agreed to participate, four centres—including that of the Assistant Principal Examiner—were selected, giving a total sample of 485 candidates. Of these participants, 451 (93.0%) sat the Higher tier paper and 34 (7.0%) sat the Foundation tier paper. This specification tends to have a high proportion of Higher tier entries relative to Foundation tier entries—67.1 *per cent* on Higher tier versus 32.9 *per cent* on Foundation tier in 2009—though clearly not as high as in the sample. All four

centres had, at that stage in their courses, concentrated on preparing their candidates for questions on Settlement and on Agriculture (the format of the paper is described under ~~"~~Materials", below). This made it very likely—though it did not guarantee—that the candidates in the sample would attempt the same questions in the mock examination, thus facilitating the analyses.

Materials

The question paper 3031/2 is divided into sections, A, B, and C. Each section contains two multipart questions and candidates are required to answer one question from each section:

Section A 1) Population ***or*** 2) Settlement

and

Section B 3) Agriculture ***or*** 4) Industry

and

Section C 5) Managing Resources ***or*** 6) Development

The participating centres were provided with the examination papers for the mock exam. For each centre, 50 *per cent* of participants were given the unmodified 3031/2 paper from June 2009, which used the conventional alphanumerical numbering system throughout. The remaining 50 *per cent* were given a modified version of the same 3031/2 paper: Section A used the conventional alphanumerical numbering system, whereas Sections B and C used the sequential numbering system. The question numbering system was the only modification made to the paper; the format and content was consistent with the unmodified paper.

Attached to each paper was a form on which candidates were asked to enter the times at which they started and finished each section of the paper, along with the numbers of the questions they attempted. For the modified papers, this form included a box in which candidates could comment on their experience of using the alternative, sequential numbering system.

Procedure

Each centre was asked to distribute the question papers so that approximately half of their participants were given a modified version, and the other half given an unmodified version. Centres with both Higher and Foundation tier candidates were asked to maintain this split within the tiers (although the data suggest that this did not happen for Foundation tier candidates). Teachers were encouraged to inform candidates before the examination that some of them would be answering a modified paper and that the modifications were made only to the question numbering system and not the question format or content.

The mock examinations were carried out under examination conditions. One hour and thirty minutes in total are allowed for the paper, with each section of the paper intended to be completed in thirty minutes. In some centres, candidates had only been prepared for Sections A and B of the paper, so were allowed one hour to complete the examination. To encourage accuracy, candidates were asked to record their start and finish times for each section as they occurred, rather than retrospectively. After the examination, candidates who sat the modified paper were able to leave comments about their experience of using the sequential numbering system.

The papers for three centres were marked by the Principal Examiner for the paper, whilst the Assistant Principal Examiner marked his own centre's scripts. The marked papers were returned to centres once the item level data had been entered at AQA.

RESULTS AND ANALYSES

Time taken

Table 1 shows the mean and standard deviation of the time taken by candidates to complete each section of the question paper by tier and group (unmodified/modified). Some centres in the sample allowed their candidates to complete the whole paper in 90 minutes, whilst others allowed them to complete Sections A and B in 60 minutes. As candidates were allowed 30 minutes per question either way, and as every centre assigned half of their candidates to the unmodified group and the other half to the modified group, this fact should not impact on the interpretation of the time data. The only things to consider, perhaps, are that the times for Section C represent fewer candidates than those for Sections A and B and that the numbers completing the Foundation tier papers are very low, so those particular figures are unlikely to paint a representative picture.

The key data here are the differences between the modified and unmodified groups on Section A and Section B of the Higher tier paper. The 20 seconds difference ($F [1, 432] = 0.328$, $p = 0.567$) between the two groups on the common Section A suggests that it is reasonable to assume that they should take a similar amount of time on subsequent sections and this is the case: in Section B, the group sitting the modified question paper were, on average, only 38 seconds faster ($F [1, 430] = 1.420$, $p = 0.234$) than those sitting the unmodified paper.

Table 1. Mean and standard deviation of the time taken by candidates per section by tier and group (unmodified/modified).

		Higher Tier			Foundation Tier		
Section	Group	N	Mean (mm:ss)	Std. Deviation (mm:ss)	N	Mean (mm:ss)	Std. Deviation (mm:ss)
A	Unmodified	213	27:16	06:03	23	23:55	06:40
	Modified	221	27:36	06:06	5	20:36	05:51
B	Unmodified	212	25:37	05:04	23	22:26	05:43
	Modified	220	24:59	05:54	5	21:48	07:57
C	Unmodified	100	28:17	04:53	10	27:54	11:13
	Modified	100	27:39	06:17	4	22:45	10:43

Performance

Figure 1 shows the relationship between scores on Section A Question 2 and scores on Section B Question 3 for the two groups of Higher tier candidates: those who sat the modified version of Section B (n=208) and those who sat the unmodified version of Section B (n=200). Section A was identical for both groups. Candidates in the modified and unmodified groups achieved comparable scores on Section B Question 3 when their scores on Section A Question 2 had been taken into account: a score on Section A predicts almost exactly the same score on Section B regardless of whether the paper was modified or not. This is borne out by the analysis of covariance, which shows a reliable relationship between scores on Questions 2 and 3 overall

(F [1, 405] = 262.696, p < 0.001) and no difference in that relationship between candidates in the modified and unmodified conditions (F [1, 405] = 0.889, p = 0.346).

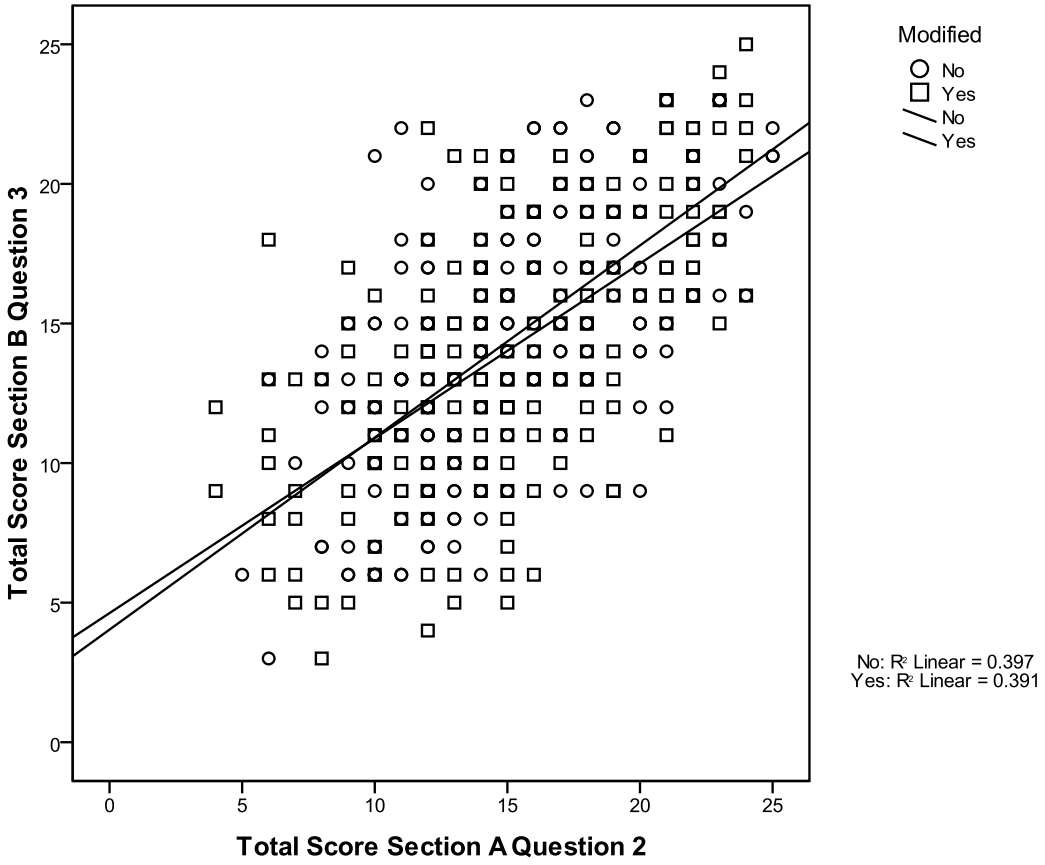

Figure 1. Section A Question 2 scores versus Section B Question 3 scores for Higher tier candidates who sat either the modified or the unmodified version of Section B

Question facilities

Although overall question score was not affected by modification of the question numbering, it is possible that some part-questions may have been affected but that any effects are masked in the overall question score. To investigate this possibility, facility indices were calculated for the part-questions, using the same sample of candidates that was used in the analysis of covariance.

It is clear from Figure 2 that the facility indices for the unmodified and modified groups are very similar for all parts of Question 2. Given that both groups sat the same unmodified version of this question, it appears that the ability profiles of these groups are similar; an independent samples t-test confirms that the overall question means of 14.92 (unmodified) and 15.25 (modified) are not significantly different (t [406] = 0.753, p = 0.452). Given this, it is reasonable to expect the facility indices for Question 3 to be very similar for the unmodified and modified papers, unless the alternative numbering systems produced between-group differences. Figure 3 shows that the facility indices for all parts of Question 3 are very similar and that the relative difficulty of items did not vary according to whether the paper was modified or unmodified.

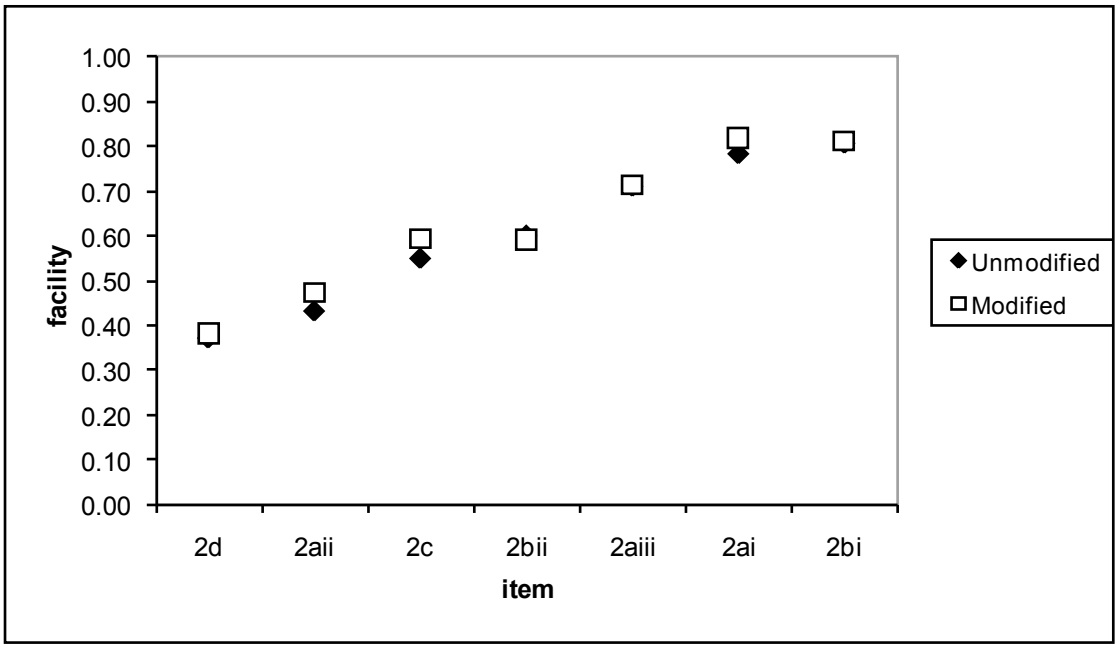

Figure 2. Facility indices for the items constituting Question 2 by candidate group

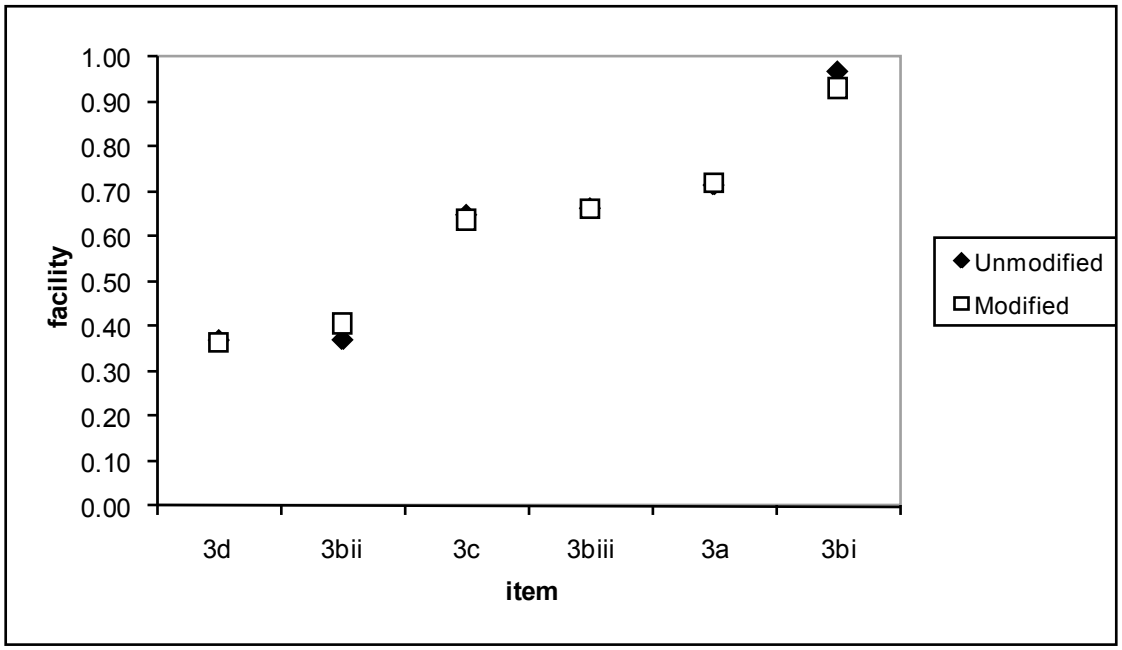

Figure 3. Facility indices for the items constituting Question 3 by candidate group

Rubric infringements: overshoots

On modified question papers, the question numbers do not signify a change of topic in the way that those on the unmodified papers do. For example, changing from Question 8 to 9, where the questions are numbered 3 to 14, does not provide the same cue as changing from Question 3(d) to 4(a), where the questions are numbered 3 to 4. For that reason, it was considered whether candidates sitting the modified papers were more likely to have accidentally ~~o~~vershot" the end of topics on questions with modified numbering than were the candidates sitting the

unmodified papers. As there are three themed sections in the paper—A, B, and C—each containing two topics, it is more likely that candidates would have overshot within sections, i.e. from Population to Settlement (Section A), Agriculture to Industry (Section B), or Managing Resources to Development (Section C), than across sections, where the themes would change more dramatically.

The most common combination of topics answered by candidates was Settlement (Section A, Q2) and Agriculture (Section B, Q3). These questions are back-to-back, and Section A used the conventional alphanumerical numbering system on both modified and unmodified papers, so most candidates were only at risk of overshooting within Section B from the end of the Agriculture question (Question 3/Questions 3 to 8) into the Industry question (Question 4/Questions 9 to 14).

Candidates were classified as overshooting if they had attempted questions in both topics within a section at least once in their exam paper. Among Higher tier candidates, only nine overshot (2.0%), but seven of these had sat the modified paper; however, of these overshooting candidates, six (one unmodified and five modified) overshot in Section A, which maintained the traditional alphanumerical question numbering style in both versions of the paper. The candidates who overshot in Section A scored relatively low marks for both topics (median marks of 7 out of 25 for Population and 9 out of 25 for Settlement[2]). It is possible that these candidates answered both topics intentionally because they did not know enough about either topic to fill half an hour and, by answering both topics, would be able to use the best mark out of the two towards their final paper mark.

Among Foundation tier candidates, five overshot (14.7%), three of whom had sat the modified paper. These three candidates overshot in Section C, but not in Section B. Their scores for the topics in Section C (median marks of 11 out of 25 for Managing Resources and 12 out of 25 for Development) are comparable to their scores for the Settlement topic in Section A (a median mark of 11 out of 25) and the Agriculture topic in Section B (a median mark of 9 out of 25). Considering this, the fact that they did not overshoot in Section B, and the fact that the overshoots which occurred were substantial (one candidate overshot by seven questions and the other two attempted all questions within both topics[3]), it seems plausible that, rather than there being ambiguity as to where the end of the first topic finished in Section C, candidates simply had time remaining at the end of the exam and used it to make a double attempt at Section C. Two candidates, who had sat the unmodified paper, substantially overshot in both Sections A and B. Their total marks for the paper were relatively low (less than half marks in each case). It seems very likely that these candidates were attempting parts of all the questions because they could not fill the available time by fully answering the intended number of questions.

There is little evidence here to suggest that the sequential numbering system led candidates to overshoot more than they might have done using the alphanumeric system. The only occasion on which candidates sitting the modified paper constituted a significant majority of those

[2] Median marks were calculated; however, some candidates were seen to have overshot to a greater extent than others. Therefore, a low mark in a topic may be due to few questions being attempted or a relatively short overshoot: it is not necessarily a true reflection of a candidate's ability on a topic. Furthermore, in our analysis, an attempted question is defined as one where a candidate wrote an answer and was awarded a mark of at least 0. Questions to which candidates gave no corresponding answers were classed as un-attempted.

[3] An overshoot of seven questions does not necessarily mean that the candidate realised they had accidentally answered too many questions and stopped after the seventh question; they may have gone through further questions in the overshot topic but were unable to give an answer to any of them. Therefore, these further questions were regarded as un-attempted.

overshooting was on Section A, where the question numbering was unmodified anyway. It seems very likely that, where candidates overshot, they did so deliberately because they wanted to attempt each topic, probably because they were unable to answer any individual topic thoroughly enough to fill the time available in the exam. Bearing in mind that, for this study, we purposefully did not change anything other than the question numbering system when modifying the papers, there remains plenty of scope to alter the formatting of sequentially numbered question papers to make it even clearer to candidates where one set of questions ends and another begins.

Candidates' comments

The candidates who sat the modified papers were asked to comment on their experiences of using the sequential numbering system and how it compared with the alphanumerical numbering system. 195 responses were received and, on reviewing the comments, the following three common themes emerged:

1. Indifference
2. Sequential numbering made answering the questions easier and quicker than alphanumerical numbering
3. Alphanumerical numbering differentiated between sections and showed relationships between questions better than sequential numbering.

In total, 76 (33.9%) of the candidates were indifferent to the type of numbering system used, 78 (34.8%) thought that sequential numbering made answering the questions easier and quicker than alphanumerical numbering, and 70 (31.3%) thought that alphanumerical numbering differentiated between sections and showed relationships between questions better than sequential numbering[4].

Some typical responses were:

> *I felt more comfortable with the new system as it was easier to see how much of the test I had completed. I definitely prefer the new system.*
>
> *The sequentially numbered system took less time than the alphanumerical numbering system and was easier.*
>
> *I didn't really notice a difference between the two. However, I prefer the alphanumerical numbering system because it gives each topic a logical progression and groups similar questions together.*
>
> *I didn't really notice a change. However, [it was] slightly harder with sequential numbering system as I didn't know which questions were linked together.*

Many candidates recorded multiple comments, which fell across the three themes. This suggests that both numbering systems have their different strengths and weaknesses, from the

[4] Where candidates gave several responses which covered different themes, each response was included. Therefore, the number of responses was greater than the number of candidates who gave them.

candidates' point of view, but that the candidates in our sample did not find one particularly preferable to the other. This is supported by the statistical analyses, which found no substantial effect of numbering system on candidates' performances. Lastly, as only the question numbering system, and not the format of the paper, was modified, some of the weaknesses of the sequential numbering system could be obviated if applied in practice. Particular weaknesses, and possible solutions to them, were mentioned by several candidates, including:

> *It was harder to tell when I should use the diagrams or graphs with the sequential numbering system.*
>
> *The size of font for subject headings e.g. Industry (Section B) should be bigger. At the end of the question set, it should say 'End of question set'.*

DISCUSSION & CONCLUSION

The purpose of this study was to determine whether there is an evidential basis for maintaining the alphanumeric question numbering system, rather than simply adopting the sequential system for all question papers. The quantitative analyses presented here suggest that the choice of system is unlikely to have any measurable impact on candidates' performances, and the candidates' comments on each system indicate no strong preference on their part for either system, each having some advantages and disadvantages compared with the other. These findings perhaps need to be considered in the context of this study and its limitations: only one subject, Geography, was included and most of the candidates were entered' for the Higher tier. Had we been able to use Maths candidates, and/or had most of them been entered' for the Foundation tier, it is possible that the results would tell a different story. However, the complete absence of any quantitative effect in this study, and the neutrality of candidates' feelings toward the choice of numbering system, suggest that this is merely a possibility, as opposed to being likely. Furthermore, candidates in this study were forewarned about the modification of the paper, but were not practiced at using it. Had an effect been found, it is possible that, with some practice, candidates could quickly overcome any disadvantage; this might also be offset by helpful modifications to the formatting of the paper.

In conclusion, the decision whether or not to discontinue using the alphanumeric question numbering system can probably be made without any concern that candidates will be disadvantaged either way. If the sequential system is adopted across all question papers, some of the advantages of the alphanumeric system could be reproduced through consideration of how the sequentially numbered question papers are formatted.

Sofia Parkinson & Neil Stringer Tuesday, 11 May 2010