STRUCTURING AND FORMATTING EXAMINATION PAPERS: EXAMINERS' VIEWS OF GOOD PRACTICE

Victoria Spalding

ABSTRACT

AQA is in the process of changing its practices, with plans to implement question banking and auto test construction, while embarking on a large scale project to improve the quality of its papers. As part of this process, Spalding (2009) reviewed the literature surrounding question paper format and structure and found that more research needed to be conducted into what constitutes good practice in structuring question papers. This report acts as an interim stage by establishing the rationale behind current AQA practice for designing question papers, and what AQA examiners consider to be good practice.

The examiners' comments fell into four categories: typesetting, presenting information, the selection of questions to include and question order. Examiners' views regarding the presentation of information and choosing questions were, in general, consistent with AQA practice and guidance from the research literature. The examiners' views of good practice only differed from AQA guidance in areas where they had received no training, such as typesetting. The examiners' views and the advice drawn from the research literature agree that ordering questions easy-to-hard is good practice. Examiners also felt that ordering questions by topic was good practice, although more research is required to determine the effectiveness of this strategy.

There are two stages in question paper construction: designing the specimen material and writing papers year on year. This report notes that the latter is constrained by the specimen paper and, therefore, future research would be best directed at specimen paper design. Several areas of good practice were identified, although there are several unknowns chiefly concerning how best to structure question papers. Consequently, two potential research projects were suggested: the ideal specimen paper design for key subjects and the impact of question banking test construction on the good practice outlined in this report.

INTRODUCTION

Research into question paper format and structure is important for several reasons. Firstly, question paper writing has been traditionally synonymous with question writing, with no reason to separate the two. However, we live in changing times with the dawn of online assessment and advancements in question banking technology. AQA has embraced question banking and plans to change its processes; future AQA papers will be constructed by a senior examiner using question banks. Secondly, there is currently a great deal of variation in paper structure across AQA subject departments and it is unclear if this variation is beneficial for candidates. There has been a call from a few members of staff within AQA to standardise the approach to question paper structure through introducing a limited number of templates for subject teams to follow, on the grounds that such a move would improve the quality of AQA's papers. Before any decision can be made, more research is required to clarify the issue. Thirdly, it is important that exam results reflect candidates' abilities and are not the result of biases associated with the nature of the assessment.

The research literature regarding font, response space, stimuli, and layout demonstrates that exam format can affect a candidate's performance. Research has shown that some typesets are better than others. Arial and Comic typeface are preferred to Times New Roman for on-screen reading and it has been argued that these fonts are more accessible to children as they resemble children's handwriting and are less formal than Times New Roman (Bernard, Chaparro, Mills, & Halcomb, 2002; 2003). The format of the provision for candidates' responses has also been shown to affect exam performance. In general, candidates achieve better marks, write longer and fuller answers, and are more precise when using combined booklets compared to separate booklets. Candidates have reported that combined booklets encouraged them to answer questions and provided some guidance regarding the length and depth of the expected response (Crisp, 2008). However, candidates have reported running out of space for long-answer questions when using combined booklets (Crisp 2008). The logical solution to this problem is to simply provide more lines; yet providing more lines may result in some candidates wasting time writing longer answers than required (Crisp, Sweiry, Ahmed, & Pollitt, 2002).

Stimuli can influence candidates' interpretation of the questions (Fisher-Hoch, Hughes, & Bramley, 1997). Images have many benefits, such as making the abstract more concrete, motivating students to respond, and making the paper seem less daunting (Crisp & Sweiry, 2006; Johnson, 2004). On the other hand, Pollitt and Ahmed (2000) found that candidates lost marks by trying to make images relevant for questions where the image was intended to provide context but was not required in the candidate's response. The location of information on the page is significant; research into memory suggests that information which is accessed first and last is most reliably recalled from short term memory (Feigenbaum & Simon, 1962; Murdock, 1962; Wiswede, Russeler, & Munte, 2007). With this in mind, important information may be best placed at the beginning, with the task instruction placed at the end. Crisp's (2008) interviewees reported that the proximity of the response space to the questions helped them respond specifically to the question asked. Having the question close to where candidates write their answer reduces the need to remember the question, which may explain why Crisp's participants found this to be helpful.

A paper's structure is a function of the type and order of the questions within it. Certain types of question are more likely to be successful at assessing some skills than others. For example, essay-based questions are historically thought to be better at assessing divergent skills, whereas multiple-choice questions are considered better at assessing breadth of subject knowledge (O'Donovan, 2005). If a paper is constructed entirely from one type of question, there is a danger that the paper will have a systematic bias. Research has demonstrated sex differences in responses to some question types. Males have been observed to do better than females in multiple-choice objective exams, while females do better than males in open-ended exams (Murphy, 1980). In some cases, gender bias may be associated legitimately with the subject content rather than question type (e.g. Maths) (Penner & Paret, 2008). Equally, multiple-choice questions have evolved to test a variety of skills and now come in many forms (Stringer, 2009a). The structure of the paper will vary with question order and there is a consistent finding in the literature that an easy-to-hard question arrangement results in better performance than a hard-to-easy question arrangement (Leary & Dorans, 1985). Anxiety may play a role in this, in that difficult questions at the beginning of an exam could heighten anxiety levels and reduce exam performance (Mulkey & O'Neil, 1999). Equally, candidates answer questions sequentially, therefore poorer candidates may run out of time, or become fatigued, and not pick up marks for easier questions placed at the end of a paper (Wilmut, 1980a).

The above points show that question paper format and structure can affect exam paper quality. However, guidance to address this issue is limited and, in most cases, does not relate directly to modern day exams. In her literature review, Spalding (2009) concluded that research should be conducted into good practice for question paper format and structure. The present paper aimed to address Spalding's (2009) recommendation by describing the rationale behind current AQA question paper design and to determine what is considered good practice. While question writing issues will naturally arise, the focus of this research was upon the format, presentation and structure of the paper as a whole (for a discussion of question writing see Chamberlain, 2009). It is hoped that the insight drawn from this research will identify opportunities for further specific research into practices that will improve question paper quality.

METHOD

Design and participants

This study had a qualitative design, consisting of semi-structured interviews and a focus group. Participants were recruited on an opportunistic basis. A total of seven examiners, all with considerable examining experience, participated in the study. Two Principal Examiners, one for GCSE Geography and one for GCSE Science, were interviewed while five members of the GCE English Language awarding committee (four Principal Examiners and the Chair) took part in the focus group. In addition, five AQA staff members were interviewed to establish current AQA practice and to provide an indication of future changes. Two members were from the question paper production department and the remaining three were subject managers.

Materials

Past papers were provided as stimulus material for the interviews with examiners. GCSE examiners received June 2007 and June 2008 subject-relevant papers from all awarding bodies. June 2008 and June 2009 AQA English Language papers were provided for the GCE focus group. The interviews and focus group each followed the same semi-structured interview schedule (see Appendix).

Procedure

The interview schedule (see Appendix) was derived from studying GCSE and A-level exam papers from AQA, Edexcel, and OCR in order to establish their constitute parts, which were grouped into those which related to format and those which related to structure. The interview schedule included formatting features including font, emboldening, numbering and images, and structural features including mix of question types, question order and linking. A loose topical approach was used, as opposed to specific questions, to allow for the discussion of issues which had not been identified by the researchers.

During the 15 minute interviews with subject managers, participants were asked to recommend experienced examiners who could be approached to take part in the second stage of interviews. The GCSE examiner interviews lasted one hour and the examiners were asked which exam paper they thought was the best overall. The GCE focus group lasted 30 minutes and the examiners in the focus group were not asked which paper they felt was the best overall as it was felt the issue was too contentious considering the papers used as stimuli were written by participants within the group.

RESULTS AND DISCUSSION

Comments from the examiner focus group and interviews were categorised according to the topics outlined in the interview schedule. The interview schedule was divided into two parts, format and structure, and analysis of the comments revealed the four sub-areas of typesetting, presenting information, questions to include and question order.

Format

The surface characteristics of papers, such as their appearance, were discussed with the examiners. Within this general area, comments fell into two sub-categories: typesetting and the presentation of information.

Typesetting

Examiners' views of good typesetting practice differed from AQA's procedures. This is perhaps unsurprising given that typesetting is controlled by AQA's paper production department and therefore examiners have little training or input in this area.

Font

AQA's typeface is currently 12pt Times New Roman, although all of AQA's papers will be in 11pt Arial font by June 2010. The move to Arial is in line with the Disability Discrimination Act which states that Arial is more accessible to visually impaired candidates. All of the participants were in favour of the move to Arial and one AQA staff member expressed the opinion that AQA papers are perceived by some to be harder, partly due to the use of the more formal Times New Roman font. There is some evidence which supports this opinion; research has found that a task is perceived as more difficult and time consuming when the instructions are printed in a hard-to-read font compared to an easy-to-read font like Arial (Song & Schwarz, 2010). However, most of the research which has been conducted compares Arial to particularly hard-to-read fonts such as Mistral; the effect may not be as strong when comparing Arial to Times New Roman. While examiners thought Arial was easier for candidates to read, they still felt that size 11pt was too small. However, increasing the font size to 12pt would make the papers longer, which was considered undesirable by AQA staff. Furthermore, the research literature suggests that the difference in readability between Arial 11 and Arial 12 is negligible (Bernard *et al.*, 2003).

Emboldening

Emboldening is standard in AQA papers and is used to emphasise the amount of responses required (give **two** examples), draw attention to negatives (which of these is **not** correct), or when referring to figures. In the examiners' views, the need to embolden should be treated on an item-by-item basis. Examiners were shown Edexcel papers in which the action word within questions is emboldened as standard. The examiners felt that whilst emboldening every action word is excessive, it should be used to highlight what is required where candidates are likely to misread the action word.

> *'I think we sometimes bold the wrong thing, and in fact I think that is something that is good* (indicating OCR paper)*, command words… We bold 'name two', then there are two marks for it so it's obviously two... So I'd rather they bold "explain" cause so many kids in Geography get "describe" and "explain" mixed up.'*
>
> Geography examiner.

> *'We are told to keep emboldening down to an absolute minimum… increasingly now we are tending now where something really is important to the sentence to*

actually embolden it and I think we should embolden more. Now that's not to say use it everywhere because if we use it everywhere it ceases to stand out to them… but… I think we ought to embolden more judiciously.'

<div align="right">Science examiner.</div>

Numbering

An alphanumerical system, sometimes referred to as 'dotties', has historically been used in AQA papers, although this is changing to a new numbering system which accommodates CMI+ marking. Most participants felt that candidates take very little notice of the numbering system and that the change in the numbering system from the old style 'dottie system' to the new numbering system will not affect candidates. However, a few felt that the numbering system provided support to the candidates by indicating links between questions.

'Candidates don't look at numbers at all, and now that in fact in terms of marking awarding it's now cause of CMI moved to item level I think we should have a debate… this fetish that we have to have a 1 and indent and so on (the dottie system) *I think that is all ridiculous.'*

<div align="right">Science examiner.</div>

'The importance of that is if it's based on a stimulus material. It's more important for the candidate now than the marker.'

<div align="right">Geography examiner.</div>

Research into the effect of the new numbering system for nested questions has recently been conducted by Parkinson and Stringer (2010). They compared performance on two versions of a GCSE Geography paper; one with an alphanumerical system, the other with the new numbering system. No difference in performance was observed between the two groups and the participants did not have a strong preference for either numbering system.

Acknowledgements

AQA uses an in-house acknowledgement system in its papers. The English A-Level examiners felt strongly that academic referencing styles, such as the Harvard style, should be used in exam papers because candidates will need to know how to reference properly when they get to university. They believed that exposing candidates to academic referencing early provides consistency and benefits their future studies. However, the examiners conceded that acknowledgements in a paper are unlikely to affect candidates' performance during the exam.

Response space

Exam papers can be presented in a booklet form, where the questions and spaces for the candidate's responses are combined, or as a separate question paper and answer booklet. The amount of space provided for an answer is normally determined by the number of marks allocated to the question and generally two lines are provided per mark. Examiners thought that the number of lines provided needs to be decided on an item-by-item basis as the number of lines provided is occasionally insufficient for some answers. Some specifications provide more lines (entitled 'extra space') in an attempt to cater for candidates with large hand writing, while not giving the impression that candidates have to fill the space. Most examiners felt that 'extra space' is a suitable solution to the problem, although the Geography examiner had reservations.

<div align="center">5</div>

> *'I don't think we should have a hard and fast rule… but what they've gone for now which I do not like is extra space… but why put extra space. Why not give them more lines?'*
>
> Geography examiner.

Crisp's (2008) research suggests that for many question types combined booklets are advantageous, although she concluded that separate booklets should be used for exams that require long answer questions, particularly for higher demand AS and A2 exams.

Presenting information
Examiners' views of good practice for presenting information were generally in agreement with AQA guidance provided in examiner training materials, and the findings drawn from the literature.

Stimuli
AQA's papers have been criticised for being too sanitised and these examiners believed that making papers engaging is an important part of the paper writing process. One examiner commented that if the specimen papers looked more attractive, this would not only improve AQA's market share, but it would also have educational benefit. Examiners felt that the use of stimuli makes a paper more interesting and using stimuli to mark the start of a new topic made the questions seem less daunting to candidates. Examiners said they also included graphs or tables as stimuli in order to assess a particular Assessment Objective (AO); for example, to assess candidates' ability to use skills and techniques appropriate to the subject. The examiners felt that stimuli must have a specific purpose, or they will act as 'noise' in an examination, and these stimuli need to be: interesting, useful to the assessment, different from other stimuli on the paper and relevant to the questions asked. This view corresponds with findings from previous research (Crisp, 2008; Crisp & Sweiry, 2006; Johnson, 2004). Additionally, one examiner described using stimuli as a springboard for ideas for writing subsequent questions. All examiners thought that colour should be used more frequently as colour makes the papers more inviting to candidates and results in a better-looking product. Examiners thought that clip art looks unprofessional and is not as appealing as original art work.

> *'We try to break it up a bit… to make it more visually appealing, I mean, yes, first and foremost what is best for the question is the main thing… but the aesthetics I think is important. I mean they are under enough stress let's make it as pleasurable as we possibly can.'*
>
> Geography examiner.

> *'What you're often looking at is a scenario which you'll get the idea from a newspaper article a magazine article or whatever. And having got this scenario you then see what AO2 skills you can tease out of it.'*
>
> Science examiner.

> '*They'll have been trained to look at this information and the teacher will say read all that because they are going to ask you about it and when they read all that and you don't ask them about it then they think, "well that was a waste of time". And when they come onto something that they really do need… then they don't bother cause they hadn't had to previously so you got to be consistent.*'
>
> Science examiner.

6

Written information

Examiners thought that rubrics should be clear, concise and easily identifiable for candidates and that text which provides information should be staggered so that candidates are not flooded with too much information. They said that the further away the information is from the question, the less likely it is that the candidate will use it. They thought that if a question is separated into sub-parts, the information for the questions should be similarly separated so that candidates do not have to search through a single large body of text.

> *'It's important that any question on the image should be next to it, and the same with information… if there is a lot of information you should very much consider splitting it up so the bits of information come with the actual questions… because what you will find invariably is that if there is some information on one page and you ask questions about it on the opposite page, they'll tend to ignore it… if it is not on the same page we should consider repeating it, like we do on-screen'.*
>
> Science examiner.

Examiners felt that if a candidate is required to draw a diagram then an empty box should be provided. In their opinion, candidates tend to look for lines and borders to find questions within combined booklets; an empty space might be interpreted as the end of a question and this could result in questions being missed. The question should be the last thing the candidates read before they respond.

> *'One of the big problems we often have is kids actually miss questions. So what they've been asked to do has got to be clear… The main thing is that what they have to do must be clear and stand out and it must be the last thing they read before they start doing it.'*
>
> Geography examiner.

Research literature into memory and recall supports the examiners' views as it has been found that information which is received last is most reliably recalled (Feigenbaum & Simon, 1962; Murdock 1962; Wiswede *et al.,* 2007).

Structure

Questions to include

Examiners' views of good practice for choosing questions were also generally in agreement with AQA guidance and the findings drawn from the literature.

Assessment Objectives

Examiners stated that a good exam paper covers all of the AOs according to their appropriate weighting, as stipulated by the Regulator. In order to have good validity, questions within an exam paper must assess knowledge and skills in the way outlined by the subject criteria and AOs. It is important to note that this is one of many requirements for producing papers with good validity (Stringer, 2009b).One examiner commented that it is often harder to write questions for some AOs than for others and suggested constructing a paper by first building questions covering the hardest AO, and then building the easier AOs around this framework. He felt that this method helps to ensure that all the AOs are covered by the desired proportion of marks.

Sampling content

Examiners believed that a good paper includes questions which sample from every area, although this does not necessarily need to be balanced within a paper. They felt that, ideally, an exam should address content in new and relevant ways. Equally, they felt that the content must differ from previous years to ensure that the papers are not overly predictable and that the whole syllabus is covered over several exam series. One examiner felt that examiners frequently have an unconscious bias towards certain areas of the syllabus. He suggested that a useful method to ensure that all areas of the syllabus are sampled equally across series is to refer to the specification outline and keep a record of the areas covered in past papers.

Variation of the questions

While questions must obviously not be repeated from previous series, examiners felt that care must also be taken to ensure that questions within a paper are not too similar. In their opinion, a question becomes redundant if it assesses the same skills and subject areas as other questions in the paper. Equally, examiners felt that in order to keep the candidates engaged, questions should be interesting, novel (within reason) and not repetitive. Examiners thought variation in types of question within a paper also makes the exam interesting for candidates. They felt that candidates will have a preferred style of question on which they perform better. They thought that as candidates differ in their preferences for question types, a paper which includes a variety of question types provides the best opportunity for every candidate to show their knowledge. The examiners' views correspond with the research on variation within papers, which suggests that varied questions types are beneficial (Murphy, 1980; O'Donovan, 2005). The examiners also believed that variation provides a fresh start for candidates after a difficult question; the candidates are less likely to be de-motivated if the next question looks different from one which they had found difficult. Examiners reported that the format of the specimen paper limited their choice of the types of questions they could use in subsequent papers.

> '*It's an intuitive feel that candidates like different kinds of questions, and even on Science A where we have objective tests as such we try and, even there, vary the types of questions within there… Individual candidates will be able to do individual things better than others, so we do feel that by providing a range of stimulus then you're giving every candidate a fair chance, at least something they can do.*'
>
> Science examiner.

Examiners thought that marks needed to be made available to candidates of all abilities for papers to discriminate effectively between candidates, and they described a good paper as one that has a good spread of high- and low-demand material. Examiners must understand the level of difficulty of their questions in order to ensure this. AQA is currently developing a new system to report on the functioning of question papers which will feed back question-level information to subject departments. Such question-level statistics may help examiners to gauge the difficulty of their questions in future series and ensure a good spread of difficulty in papers. Some examiners felt that including too many one-mark questions, which do not allow candidates to show their knowledge, reduced the quality of the paper. One examiner suggested that multiple-choice questions can be used to make a paper more accessible to low ability candidates.

Question order

Easy-to-hard arrangement

Examiners felt that the first few questions in a paper should be relatively easy so that candidates can relax and not become over anxious. It was their opinion that an easy-to-hard arrangement ensures that candidates are not discouraged early on, thus avoiding the possibility

of candidates failing to attempt questions. Additionally, they believed that placing the easier questions towards the beginning of the paper ensures that less able candidates attempt questions on which they can achieve marks before they run out of time. The easy-to-hard ordering of questions has become an unwritten rule in AQA and most AQA papers have an incline of difficulty. Happily, the examiners' views and the advice drawn from the research literature agree that ordering questions easy-to-hard is good practice (Leary & Dorans, 1985).

> *'There is a place for one or two questions at the start of the paper, to settle candidates in, which have fairly high facilities... if you get the paper right, totally right, obviously there will be some bumps, but it will start up there* (high facility) *and it will end hopefully not lower then 30% or something like that, so that you've got that gradient through the paper.'*
>
> Science examiner.

In contrast, the GCE English Language examiners considered that using an incline of difficulty was inappropriate for their paper as it consisted only of a few long answer questions of equal difficulty. These examiners did feel that having a few easy questions to start with would be a good thing but due to the restrictions associated with optional questions they were unable to include short starter questions.

Coherence

In the examiners' opinion, the order of the questions affects the coherence of a paper. They thought that questions should naturally follow on from each other, particularly if they are grouped into sub-parts, and that they should make sense when linked together. Examiners felt that grouping questions by topic or theme reflects how candidates acquire their knowledge. In their experience, candidates tend to compartmentalise their learning, particularly at GCSE level. If questions jump across topics, candidates are forced to think synoptically. The examiners felt that this is very demanding and should be avoided, unless the paper is specifically designed to assess synoptic learning. Whilst examiners felt that ordering questions by topic was good practice, more research is required to determine the effectiveness of this strategy.

Continuing question order from previous years

Both unpredictable and over-predictable question papers have poor validity as both place too much of a premium on exam technique and preparation rather than being fair tests of the required knowledge, understanding and skills (Alton, 2008). A balance needs to be found between exam papers which are familiar and those which are over-predictable. The Regulator states that exam papers should be comparable to past papers, although AQA provides little formal advice on how to structure exam papers. In terms of an exam paper following the specimen paper, AQA advice is vague and its interpretation varies between subject departments. Whilst the ambiguity of AQA's advice does lead to some confusion, instruction needs to be vague in order to address such a wide range of papers.

On the whole, examiners felt that it was good practice to continue a similar question order to the specimen material as they felt this made the paper familiar to candidates. Examiners felt that it is beneficial to follow the same order of questions year on year for specifications which include inserts and additional reading material. They reasoned that candidates become familiar with the structure and can easily navigate through it. The English Language A-level examiners said that their papers are particularly well suited to a stable question order as they include a large amount of source material. Some examiners, however, felt that for papers where navigation is not an issue, limiting question order to that of previous years may not be of benefit.

Some examiners struggled to consider ways of structuring papers which differed from the specimen material. One examiner expressed a desire to have more freedom but could not describe how he would change the structure of the paper if he was at liberty to do so.

GENERAL DISCUSSION

The aim of this paper was to address the first stage of research suggested by Spalding (2009, p. 11), who stated that *'AQA will need to access the current good practice and processes for question paper construction for a variety of paper types'*. This report has described the rationale behind how AQA currently designs question papers and has outlined what AQA examiners consider to be good practice. Exam paper format and structure were investigated and the advice that can be drawn from this paper has the potential to impact AQA processes at two levels: specification paper design and the subsequent writing of exam papers.

Examiners' views regarding the presentation of information and the choosing of questions are, in general, consistent with the training that they have received from AQA. The examiners' views of good practice only differed from AQA guidance in areas where they had received no training, such as typesetting, or where AQA do not provide guidance, such as the ordering of questions. Examiners' views are prone to a positive bias, in that they believe their own approach to be best practice, which is understandable given their experience and AQA training. In other words, their views of good practice correspond tightly with their own procedures, which naturally follow AQA guidance and training. Given this bias, examiners' views cannot be taken as comprehensive objective evidence for good practice and more research needs to be conducted in several areas.

Discussions with examiners revealed some insight into how they construct papers. One examiner had a very clear idea of how he constructs his papers. His first step is to find a range of stimuli material which relates to the subject content. He then selects the most appropriate stimuli, based upon the required AOs, before writing questions around this material. He felt that this method of paper construction produces coherent and discriminating papers by establishing the correct level of difficulty whilst ensuring that the appropriate subject content and AOs are covered. In contrast, other examiners were unable to articulate how they constructed papers. One examiner commented that while he obviously had a way of going about it, he was not conscious of how exactly he wrote his papers. Whilst this lack of awareness does not reflect a lack of ability, it is problematic for AQA as it is much harder to develop audit trails for paper development and pass on expertise to inexperienced examiners. Whitehouse, He and Wheadon (2008) found that while on the whole examiners felt there were considerable benefits to using question banking test construction, they did feel limited as they could not modify questions to fit their paper. It would be useful to know more about how examiners construct papers and how question banking may affect this process.

The GCSE examiners were asked which paper they felt was best overall and the reasons behind their decision. Both examiners felt that the mix of question types and spread of demand played into their decision. Some papers were ruled out as they had one particularly bad question and others were ruled out as they had too many questions of the same type, but the deciding factor was the quality of the questions within the paper. While format and structure do affect the quality of a paper, the quality of the individual questions is still pivotal in ensuring the production of good papers. Any evaluation of the format and structure of exam papers needs to take into account that the quality of a paper is dependent upon the quality of the questions within it.

This report has highlighted that a great deal of question paper format and structure is pre-determined and possibly constricted by the specification and the specimen paper. Awarding bodies have a restricted amount of time to produce specimen material; for example, in 2006 the awarding bodies had only three months to produce the new syllabuses and assessments (Baird & Lee-Kelley, 2010). As a result of this short time frame, less time is spent producing the specimen papers than the subsequent exam papers. Specimen papers are not normally trialled and do not go through the usual question paper evaluation committee (QPEC) meeting. The examiners expressed frustration with the time limits surrounding specimen paper development, and felt that the quality of the specimen papers was often compromised as a result. Several aspects of question paper format and structure, in particular, can only be addressed during the design of the specimen paper. Therefore, this report recommends that future research would be best directed at specimen paper design. Further, the current study is limited by only involving examiners from a small number of subjects. It is likely, for example, that the ideal structure for a foreign language paper will differ from the ideal structure for a Maths paper, and so the advice for good practice will vary accordingly. More research is required to fully understand subject-specific effects.

CONCLUSION

Current AQA advice regarding question paper format is, in general, considered good practice by examiners and has corresponding support from the research literature. On the other hand, there is limited guidance available for question paper structure from AQA or the research literature. Although some views of good practice have been described in the current study, more research is required to produce guidance for designing question papers. Bearing in mind the findings from the current study and AQA's future plans to use question banking and auto test construction, this report suggests two avenues for future research.

Firstly, a methodology for trialling specimen papers should be derived. This would be beneficial given that the specimen paper dictates the structure of the following exam paper, hence impacting on the quality of subsequent papers. Due to the restrictions from the Regulator, specimen papers are currently produced in a very short time frame and such research would be subject to a longer time frame being made available. How the specimen paper is best structured should be investigated and it would be sensible to investigate subject-specific effects. For example, is there an ideal structure for a Science exam paper and, if so, what is that structure? Would this structure also be ideal for Geography? If standardised templates were to be used, which structures should they be based upon and how many templates would you need to cover all the exams that AQA offers? The methodology would require objective measures of paper performance alongside feedback from candidates and centres in order to avoid any bias towards current AQA practices. Such research would also need to take into account the quality of the individual questions in the paper. AQA is currently developing question paper functioning reports which may act as an appropriate objective measure of performance of question papers.

Secondly, it would be beneficial for AQA to investigate the impact of using question banks on paper construction. It is possible that using question banks in a semi-automated process will dramatically change the way in which examiners construct their papers. It would be interesting to discover if the rules of good practice outlined above could be achieved effectively using question banking.

While the current study produced modest recommendations, this report has highlighted some knowns and unknowns regarding question paper format and structure. Two areas of research

11

have been suggested which will result in a deeper understanding of AQA current processes and bring about continued improvement in the quality of AQA papers.

RECOMMENDATIONS

- Further research into the ideal specimen paper design (such research should take into account subject-specific effects).
- Further research should be undertaken to investigate the impact of using question banks on paper construction.

REFERENCES

Alton, A. (2008). *Ofqual Predictability Studies Report: A study of GCSE and GCE level Examinations.* Retrieved 25 January 2010 from http://www.ofqual.gov.uk/1984.aspx

Baird, J., & Lee-Kelley, L. (2010). The dearth of managerialism in implementation of national examinations policy. *Journal of Education Policy, 24(1),* 55-81.

Bernard, M. L., Chaparro, B. S., Mills, M. M., & Halcomb, C. G. (2002). Examining children's reading performance and preference for different computer-displayed text. *Behaviour and Information Technology, 21(2)*, 87-96.

Bernard, M. L., Chaparro, B. S., Mills, M. M., & Halcomb, C. G. (2003). Comparing the effects of text size and format on the readability of computer-displayed Times New Roman and Arial text. *International Journal of Human-Computer Studies, 59*, 823-835.

Chamberlain, S. (2009). *A guide to writing examination questions: theoretical approaches and practical solutions.* Manchester: Assessment and Qualifications Alliance.

Crisp, V. (2008). Improving students' capacity to show their knowledge, understanding and skills in exams by using combined question and answer papers. *Research Papers in Education, 23(1)*, 69-84.

Crisp, V., & Sweiry, E. (2006). Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research, 28 (2),* 139-154.

Crisp, V., Sweiry, E., Ahmed, A., & Pollitt, A. (2002, September). *Tales of the expected: the influence of students' expectations on question validity and implications for writing exam questions.* Paper presented at the British Educational Research Association Conference, Exeter.

Eason, S. (2006). *GCE Geography Specification Development Questionnaire Survey, Analysis of Responses to Closed Questions.* Guildford: Assessment and Qualifications Alliance.

Fisher-Hoch, H., Hughes, S., & Bramley, T. (1997, September). *What makes GCSE exam questions difficult?: Outcomes of manipulating difficulty of GCSE questions.* Paper presented at the British Educational Research Association Annual Conference, York.

Feigenbaum, E. A., & Simon, H. A. (1962). A theory of the serial position effect. *British Journal of Psychology, 53 (3),* 307-320.

Johnson, M. (2004). This one is more me! What children think about writing test stimuli involving choice. *Literacy, 38 (2),* 90-96.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55,* 387-413.

Mulkey, J. R., & O'Neil, H. F. Jr. (1999). The effect of test item format on self-efficacy and worry during a high-stakes computer-based certification examination. *Computers in Human Behaviour, 15,* 495-509.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64 (5),* 482-488.

Murphy, R. (1980). *Sex differences in objective test performance.* Guildford: The Associated Examining Board.

O' Donovan, N. (2005). There are no wrong answers: an investigation into the assessment of candidates responses to essay-based examinations. *Oxford Review of Education, 31 (3),* 395-422.

Parkinson, S., & Stringer, N. (2010). *Numbering nested questions.* Guildford: Assessment and Qualifications Alliance.

Penner, A. M., & Paret, M. (2008). Gender differences in mathematics achievement : Exploring the early grades and the extremes. *Social Science Research, 37*, 239-253.

Pollitt, A., & Ahmed, A. (2000, September). *Comprehension failures in educational assessment.* Paper presented at the European Conference on Educational Research, Edinburgh.

Taylor, R. (2009).*GCSE English A (3702): Some insight gained from the Rasch model.* Manchester: Assessment and Qualifications Alliance.

Whitehouse, C., He, C., & Wheadon, C. B. (2008). *Item banking with a test construction interface: an evaluation of a prototype.* Guildford: Assessment and Qualifications Alliance.

Willmott, A. S., & Hall, C. G. W. (1975). O-level examined: The effect of question choice. Schools Council Research Studies. London: Macmillan Education Ltd.

Wilmut, J. (1980). *Candidates' question choice in the June 1980 Religious studies syllabus II (Multi-faith), Paper 1.* Guildford: The Associated Examining Board.

Wiswede, D., Russeler, J., & Munte, T. F. (2007). Serial position effects in free memory recall— An ERP-study. *Biological Psychology, 75,*185–193.

Song, H. & Schwarz, N. (2010). If it's easy to read, it's easy to do, pretty, good, and true*. The Psychologist, 23(2),* 108-111.

Spalding, V. S. (2009). *Is an exam paper greater than the sum of its parts? A literature review of question paper structure and presentation.* Guildford: Assessment and Qualifications Alliance.

Stringer, N. (2009a). *Response to Ofqual Science Report.* Guildford: Assessment and Qualifications Alliance.

Stringer, N. (2009b). *Assessment expertise project: validity of assessment.* Guildford: Assessment and Qualifications Alliance.

Victoria Spalding
February 5, 2010

APPENDIX:

Interview schedule

1. Format and Presentation
Font:
Examples:
- AQA and Edexcel: Times new Roman
- OCR: Arial

Questions:
- Do you think font matters?
- What is the best style?
- What do you think is the best size for font?

Rubric & instructions:
Examples:
- AQA and OCR have instructions at front in bullets.
- Edexcel have instructions at front as paragraph.
- Text information is presented in boxes.
- Information is in question.
- Placing after question.
- Within the text.

Questions:
- Who has the clearest instructions?
- Which presentation is best?
- Do you think we can make it better?
- Why do we put text in boxes?
- Do you think there is a better way of presenting supporting information?
- Do we have standard ways of presenting this information?
- Do you have a standard way of placing rubric?
- Would a standard be appropriate?
- Do you think we can improve how we present rubric?

Emboldening:
Examples:
- AQA highlights things like number of responses.
- OCR highlights action words.

Questions:
- Why do we embolden words?
- When do you think it's appropriate?
- Do you think OCR's approach is effective?

Numbering:
Examples:
- AQA format: 1 a ii, 1 a iii etc.
- Edexcel format: 1 a i, ii
- OCR format: A1 a i, ii but also A1, A2, B3, B4 etc.

Questions:
- Why do we number the way we do?
- Which do you think is best?
- Is there a better way of numbering?
- Does numbering matter?

Space:
Examples:
- Combined paper vs use separate answer booklet:
  AQA, Edexcel, and OCR use combined on some/all of their papers,
  OCR use separate.
- Amount of space provided for marks:
  AQA does 'extra space', others don't.
- Structured response.

Questions:
- Do you consider the amount of space important?
- Does the space provided reflect the answer you expect?
- Why do we provide 'extra space' on some questions? Does this work?
- Why do we structure answers?
- Do you think this is good practice?

Mark allocated:

Examples:
- All papers provide mark allocated.
- Some are specific to each question part.
- Others just do total for each question.

Questions:
- Why do we provide mark information?
- Do you think it helps?
- Can it mislead students?
- Which way is best?
- Can we do better?

Images:

Examples:
- In booklet vs on paper - frequently candidates have to refer to a booklet.
- Positioning in question: above, opposite, above and below.
- Type of image.

Questions:
- When do you use images?
- What are the benefits of images?
- When and why do we place some images in a booklet?
- Where is the best place to place an image?
- Would if be of benefit to put images in standard places?

2. Structure

Linking:
- Why do we ask questions in sections by topic?
- Do you link questions?
- Does the answer to one question lead onto the answer of the next? If so, why do we do this?

Mix of item types:
- How does an examiner balance Assessment Objectives?
- Do you try to include a mix of response types? If so, why?
- Do you try to include a mix of stimuli types? If so, why?
- How does an examiner sample content?
- Do you try to cover more than one topic area in a question? If so, why?
- Do you try to include a mix of question types?
- What are the draw backs and benefits of different question types?

Question order:
- How does an examiner achieve coherence?
- Do you try to cover more than one topic area in a question? If so, why?
- What are the factors you consider when ordering your questions in a paper (specimen papers and those which follow)?
- Do you start with an easy question? If so, why?
- Does order of questions affect flow?
- In general, is there a right way to order questions?
- What should examiners try to avoid when ordering questions?

3. Anything I have not thought of?

4. Which paper do you think is best?

5. Why do you think this paper is best?