CAMBRIDGE ASSESSMENT

An empirical exploration of human judgement in the marking of school examinations

Jackie Greatorex and W. M. Irenka Suto,
Research Division
Cambridge Assessment

Paper presented at the International Association for Educational Assessment Conference, Singapore, 21st to 26th May 2006.

Address for correspondence
Jackie Greatorex
Research Division
Assessment Research and Development
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG
UK
E-mail: greatorex.j@cambridgeassessment.org.uk
Telephone: 44 (0)1223 553835
Fax: 44 (0)1223 552700

www.cambridgeassessment.org.uk

Cambridge Assessment

Cambridge Assessment is the brand name of the University of Cambridge Local Examinations Syndicate, a department of the University of Cambridge. Cambridge Assessment is a not-for-profit organisation.

Abstract

A major theme of our recent research has been the nature and use of human judgement in the marking of school examinations. In an era of innovation and rapid development, it is important to have an understanding of these psychological processes, which have the potential to impact upon modernisation. In this paper, we present an overview of our studies in this area.

Working within a popular cognitive psychological paradigm, we explored examiners' judgements in a number of marking contexts. Both experienced and newer examiners participated in the research, in which a 'think aloud' method was utilised. GCSE and A-level examinations were marked, and both paper-based and computer-based marking formats were investigated.

We identified five distinct cognitive marking strategies, which were used in all of the contexts considered. Subsequently, a quantitative analysis of strategy usage in the traditional paper format was conducted. We will conclude this paper with a discussion of the potential implications of this research for future examination marking.

Introduction

A-levels and GCSEs play a crucial role in secondary education throughout England and Wales, and the process of marking them, which entails extensive human judgement, is a key determinant in the futures of many eighteen and sixteen-year-olds. The judgement and decision-making processes involved in the marking of some other kinds of examinations have received some serious consideration among researchers (including Cumming, 1990; Vaughan, 1992; Milanovic *et al*., 1996; Laming, 1990, 2004; Webster *et al*., 2000; Yorke *et al*., 2000). Furthermore, Sanderson (2001) has investigated the cognitive processes used by A-level examiners when marking essays in Sociology and Law. However, the judgements made whilst marking the shorter answer questions that comprise significant proportions of some GCSE and A-level examinations have yet to be explored in detail. In an era of innovation and change, it is important to understand such judgement processes and their potential to impact on modernisation. They have therefore become a major theme of our recent research.

Over the past year we have conducted some inter-related studies, the key aspects of which are the focus of the present paper. In our first study, the main aims were to identify and investigate some of the cognitive strategies used when marking GCSEs and to interpret them within the context of established psychological theories of human judgement. We begin this paper by summarising this first study, which is described elsewhere, both in full (Suto and Greatorex, *in press, a*) and in outline (Suto and Greatorex, *in press, b*). In the remainder of this paper, we move on to describe how we developed our ideas in a second study. We considered whether the same marking strategies are used (i) in A-level as well as GCSE marking; (ii) by different types of markers; (iii) in on screen marking as well as in traditional paper-based marking.

Study 1: Identifying cognitive strategies used in GCSE marking

Background

GCSE examination marking is a diverse activity, encompassing a wide range of subjects with a variety of question styles and mark schemes. It is likely, therefore, that at least some aspects of it will have parallels with some of the activities already scrutinised by judgement researchers in other contexts. Psychologists have constructed multiple models of judgement and decision-making, which have yet to be applied to examination marking, and one potentially useful theoretical approach is that of dual processing. Such models distinguish two qualitatively different but

concurrently active systems of cognitive operations: *System 1* thought processes, which are quick and associative, and *System 2* thought processes, which are slow and rule-governed (Kahneman and Frederick, 2002; Stanovich and West, 2002).

The 'intuitive' judgments of System 1 are described as automatic, effortless, skilled actions, comprising opaque thought processes, which occur in parallel and so rapidly that they can be difficult to elucidate (Kahneman and Frederick, 2002). System 2 judgments, in contrast, have been termed 'reflective', and the thought processes they comprise are characterised as slow, serial, controlled, and effortful rule applications, of which the thinker is self-aware (*ibid*. 2002). According to Kahneman and Frederick (2002), as an individual acquires proficiency and skill at a particular activity, complex cognitive operations may migrate from System 2 to System 1. For example, chess masters can develop sufficient expertise to perceive the strength of a chess position instantly, as pattern-matching replaces effortful serial processing.

There may be question types, or stages of marking, that involve System 1 processing; at times, simple and repetitive matching of a candidate's single-word response with the model answer given in the mark scheme may be all that is required. At other times, examiners might be engaged in System 2 processing; for example, when carefully applying the complex guidelines of a mark scheme to a candidate's uniquely worded answer. As examiners become more familiar with a particular examination paper and mark scheme, or more experienced at marking in general, some sophisticated thought processes may be transferred from System 2 to System 1, while others remain exclusive to System 2.

Materials and Methods

To explore the possibility of applying this theoretical approach to GCSE marking, we investigated two contrasting examinations (administered by Oxford, Cambridge and RSA Examinations (OCR) in 2004) in our study: an intermediate tier Mathematics paper, with a 'points-based' marking scheme, and a foundation tier Business Studies paper, with a 'levels-based' scheme. For both examinations, candidates' scripts comprised individual booklets containing subdivided questions with answer spaces allocated to each question part.

For each subject, a group of six 'expert' examiners (either teachers or retired teachers, with considerable marking experience), comprising one Principal Examiner and five Assistant Examiners, participated in the study. After some silent marking to

familiarise themselves with the question paper and mark scheme, and after receiving some feedback on their marking, the examiners were asked to ‘think aloud’ whilst marking identical script samples. Subsequently, they were interviewed about their marking.

Qualitative Analysis and Findings

A qualitative analysis and interpretation of the verbal protocol and interview data was conducted, in which the mark scheme and scripts were also utilised. It enabled us to propose a tentative model of marking, which includes five distinct cognitive marking strategies: *matching*, *scanning*, *evaluating*, *scrutinising*, and *no response*. An overview of the model is given in Figure 1, and the strategies are presented individually in Figures 2 to 6. Figure 7 contains a key to the other figures. We used the model to code all of the verbal protocol data according to the strategy/strategies used to mark each question part. The strategies were broadly validated not only in the retrospective interviews with the examiners who participated in the study, but also by other senior Mathematics and Business Studies examiners. However, the model is unlikely to be exhaustive, and further data might yield additional strategies for marking short-answer GCSE questions.

Figure 1 *Model summarising the processes entailed in marking a GCSE examination*

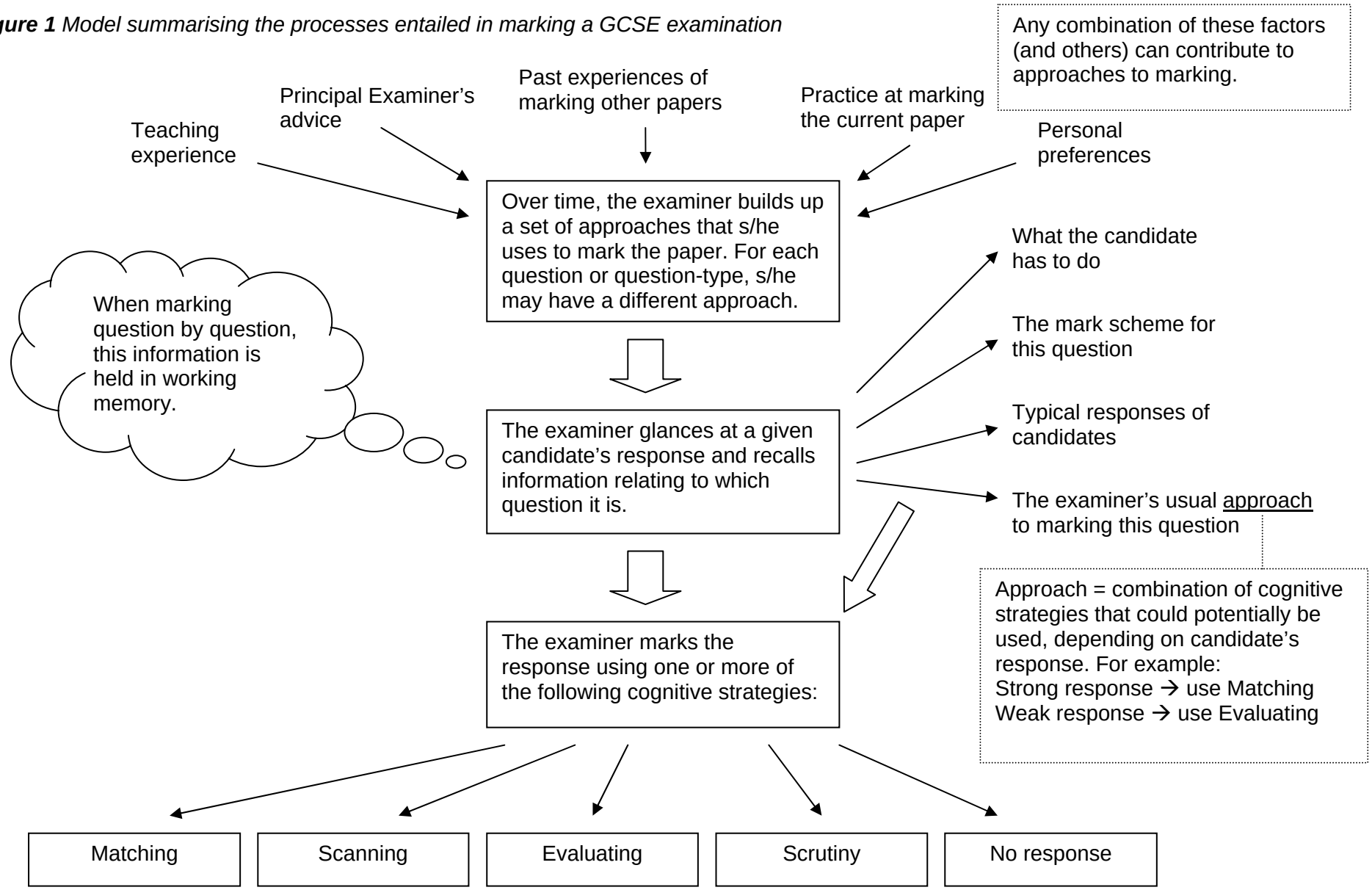


Figure 2 *The 'Matching' strategy*

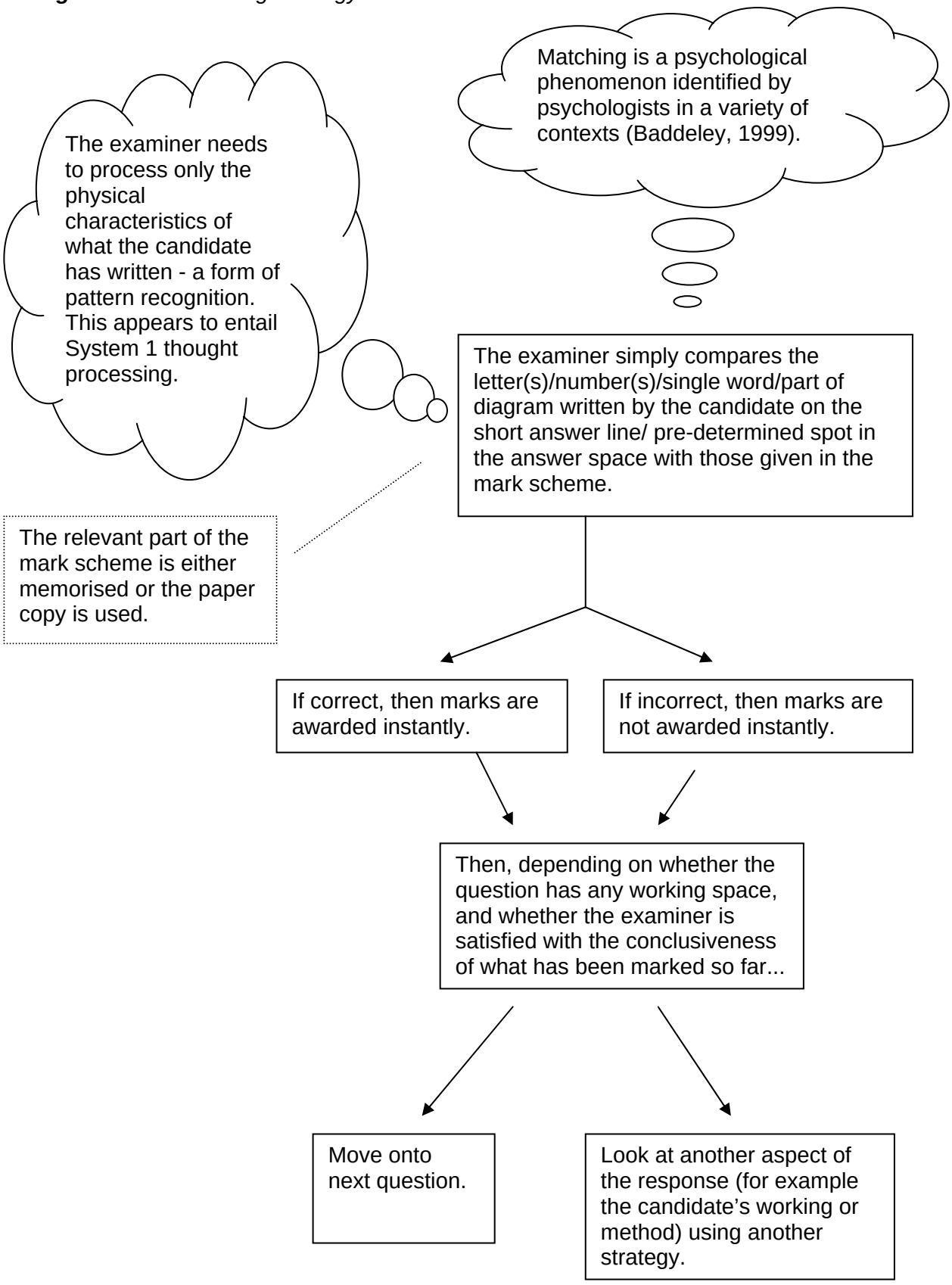


Figure 3 *The 'Scanning' strategy*

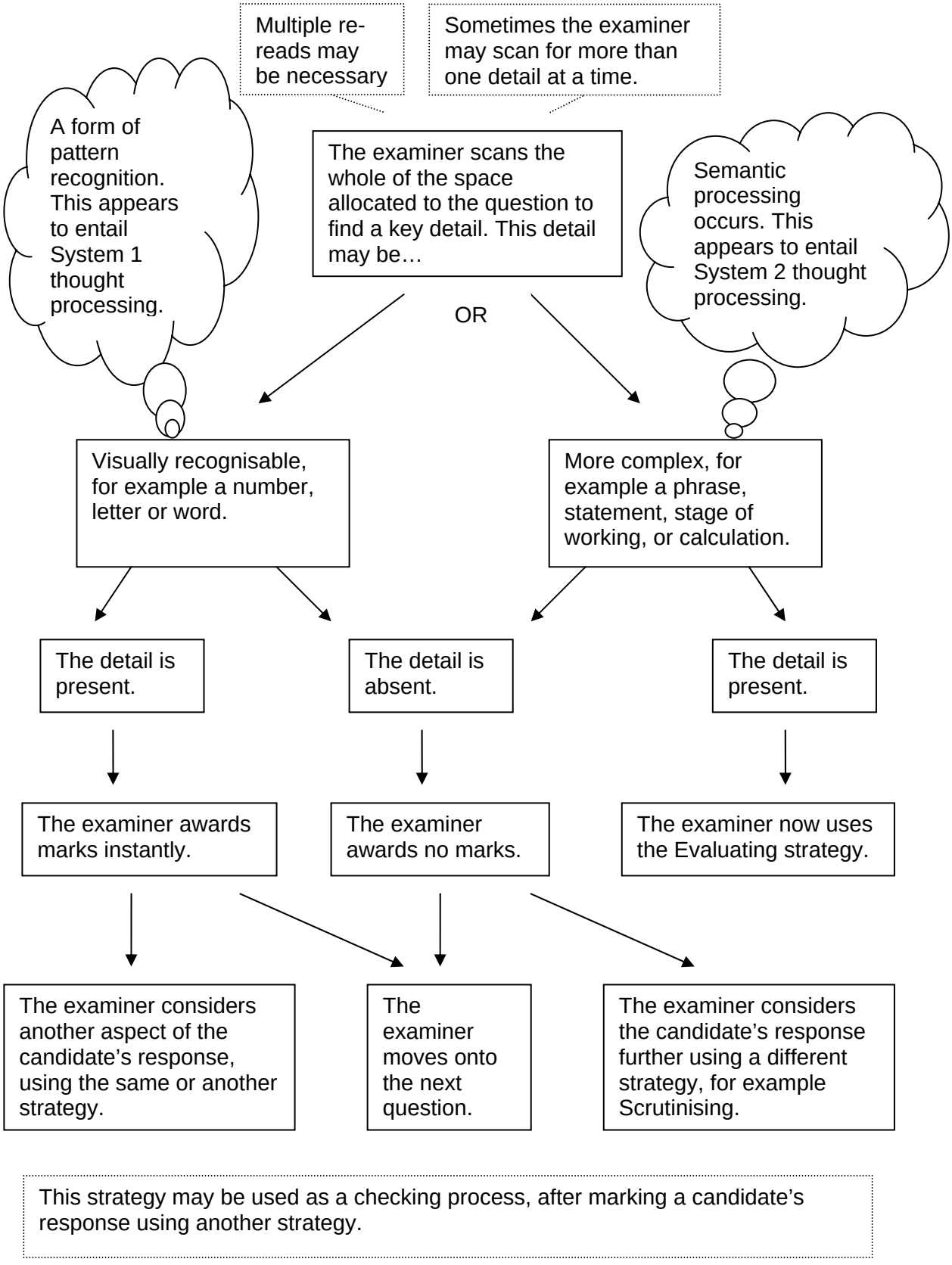


Figure 4 *The 'Evaluating' strategy*

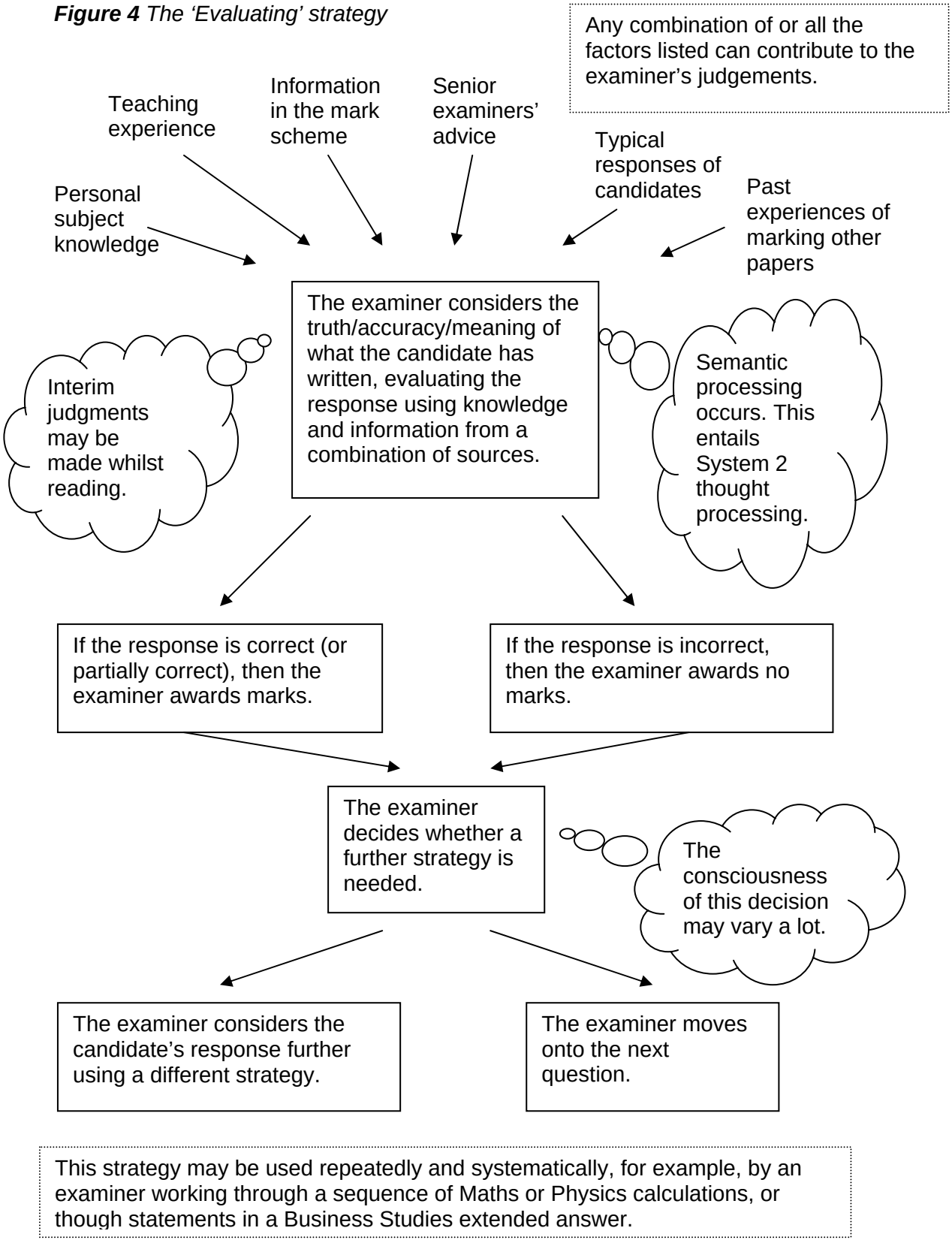


Figure 5 *The 'Scrutinising' strategy*

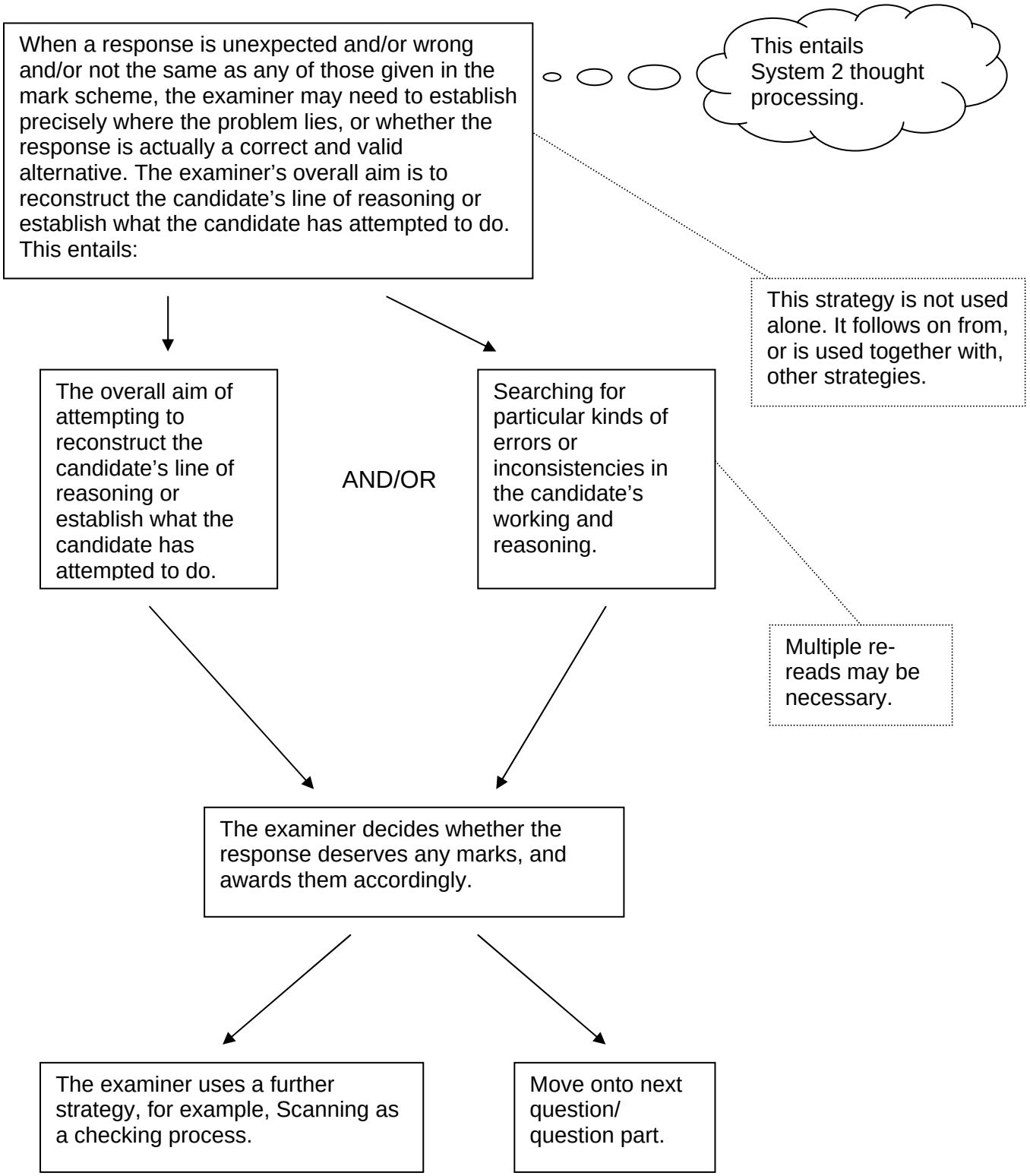


Figure 6 *The 'No response' strategy*

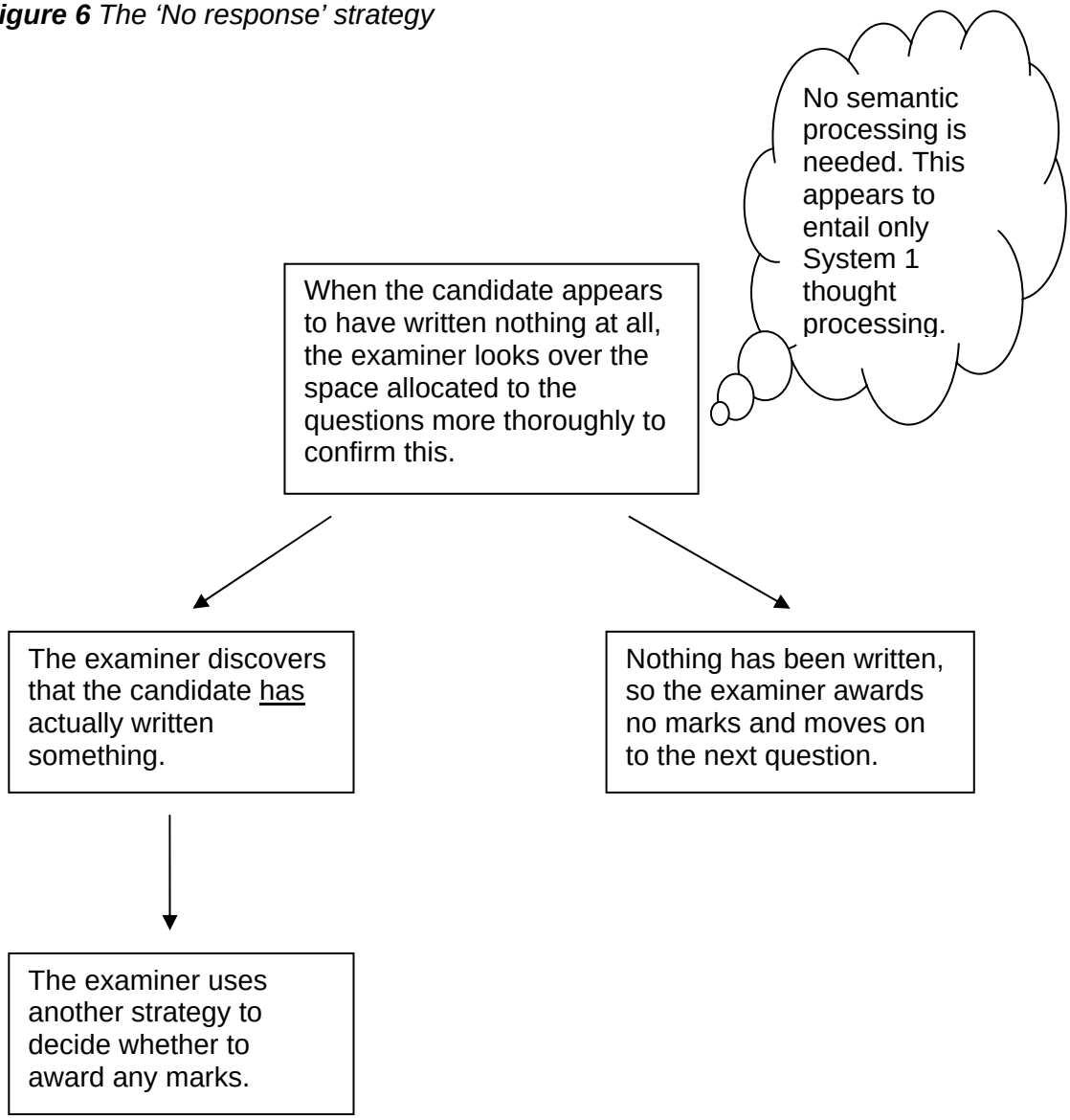
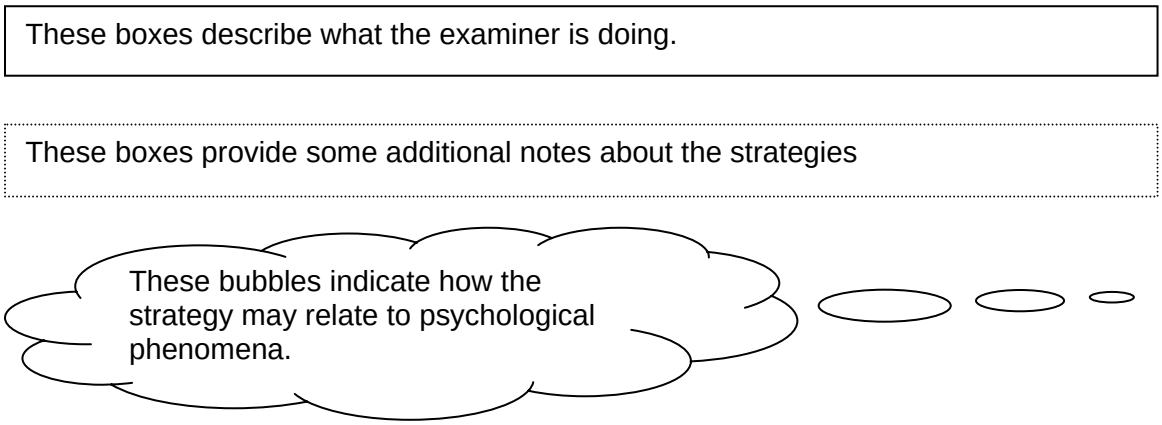


Figure 7 *Key to Figures 1 to 6*



Brief Discussion Point

In his work on essay marking, Sanderson (2001) has been careful to use a theoretical framework in which examining is viewed as a socially constructed activity entailing both explicit and implicit knowledge drawn from examiners' cultural experiences. He argues that examining can be seen as a series of (i) cognitive processes, social judgements and the quantification of judgement, and/or (ii) a form of problem-solving requiring examiners to adopt procedural strategies to achieve an outcome. In his model, Sanderson attempts to specify how these different views of examining interact. He suggests that there are two forms of difference in examining A-level essays; (i) cultural or discursive differences arising from membership of communities of practice, and (ii) individual differences based on cognitive capacity or strategic choices. Our study focused on the processes occurring while examiners mark individually and we used a cognitive psychological approach, which emphasises the individual's cognitive activity. However, this approach does not ignore the social context of assessment entirely; for example, some of the social aspects of marking are indicated in Figures 1 and 4. In particular, teaching experience, senior examiners' advice and subject knowledge are all developed through being a member of a community of practice. It is beyond the scope of our research to address the emotional aspects of marking and how they affect judgements, if at all, but this would provide another interesting angle from which to consider this topic.

Quantitative Analysis and Findings

A quantitative analysis of the verbal protocols was also conducted (Suto and Greatorex, *in submission*). Using the coding constructed for the qualitative verbal protocol analysis, we quantified the frequencies of cognitive strategy usage on each question part and for each individual examiner, as well as for all examiners marking each subject. The reliability of each individual examiner's marking was also calculated, using data from the silent marking at the beginning of the study. For each examiner, these 'experimental' marks were compared with (i) the marks awarded when the same scripts were marked professionally, the previous year; and (ii) the Principal Examiner's 'experimental' marks. Individual question parts on which significant differences in marking occurred were also identified.

There were some differences in strategy usage among individual examiners within subjects. However, the more prominent differences were between subjects and among questions. Figures 8 and 9 summarise the frequencies of strategy usage for the two subjects when all examiners (for

that subject) are considered together. They show that the Business Studies examiners used the evaluating strategy relatively more often than the Mathematics examiners, and used the matching strategy relatively less often. This may reflect the more subjective judgements involved in marking Business Studies using a level based mark scheme. Within each subject, no clear relationships between strategy usage and marking reliability were found, suggesting multiple successful ways of marking some questions.

Figure 8

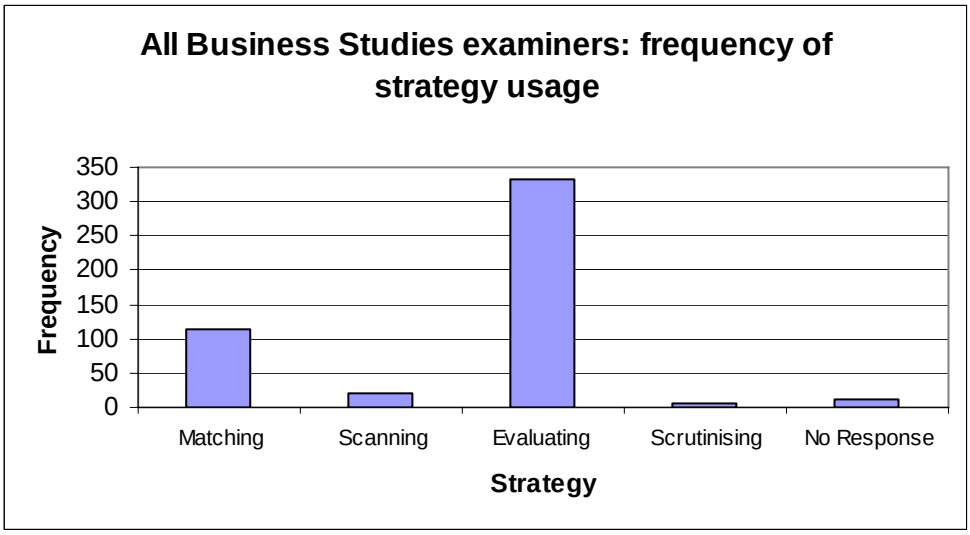
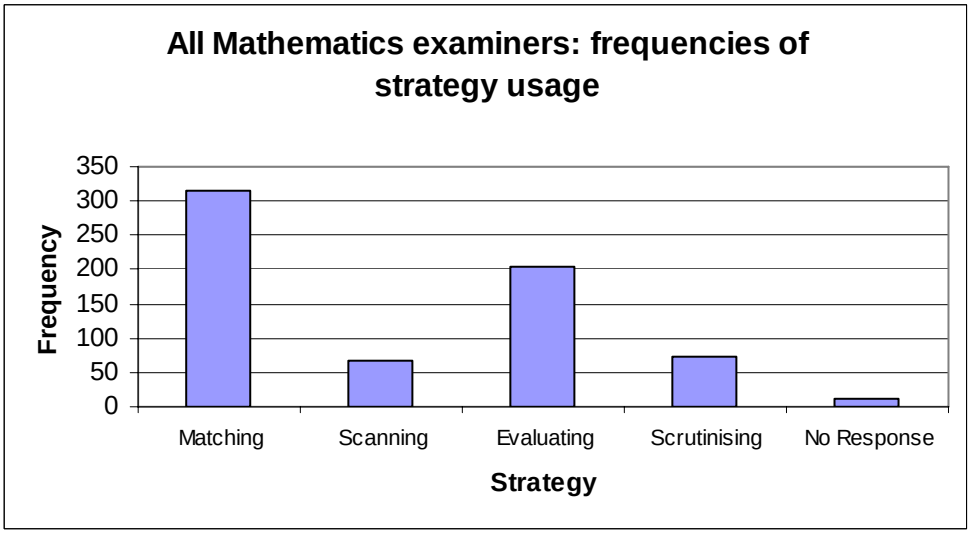


Figure 9



Discussion

Extensive discussions of this study are given in Suto and Greatorex (*in press a, b; in submission*).

Study 2: Are the five cognitive marking strategies used in other contexts?

Background

An important factor that could potentially determine strategy usage is that of marking experience. Weigle (1999) cites several studies (Huot, 1988; Cumming, 1990; Shohamy *et al*, 1992; Weigle, 1994) in which differences between the severity of grading and the rating strategies of 'novice' and 'expert' markers have been found. Given this literature, an aim of our second study was to establish whether or not the marking strategies identified among 'expert' markers in our first study are also used by examiners with less experience. Additional aims were to establish whether the same marking strategies are used in A-level as well as GCSE marking, and in on-screen marking as well as in traditional paper-based marking.

Materials and Methods

The study was opportunistic, in that we obtained verbal protocol data for reanalysis from an earlier operational development project with other objectives. Not only did this data comprise verbal protocols from both expert and subject markers[1], but there were some other key differences with the data collected in the first study. Whereas in the first study, whole scripts were marked from GCSE Mathematics and Business Studies papers, the second study's data related in part to the same GCSE Mathematics paper but also to an A-level Physics paper (also administered by OCR in 2004) which had yet to be used in research of this kind. Secondly, marking was conducted 'item by item' rather than 'candidate by candidate'. Thirdly, the second study's data related to 'on-screen marking' (or 'marking from image'), whereas the first study's examiners marked candidates' scripts in the traditional paper format. We decided that although this opportunistic data was far from comprehensive, it would provide a useful means of validating our five identified cognitive strategies and considering their relevance to this different mode of marking.

For each of the two examination papers, random samples of candidates' responses to individual questions (from the previous year's 'live' examination) were selected for marking. Each sample comprised approximately five responses to each question in a short series of questions. Due to the design of the original operational development project, slightly different combinations of questions were allocated to different markers in the study.

The categorisations of the 14 markers who participated in the study involved are presented in Table 1. All but one of the markers (a male 'expert' Mathematics marker) were new to research of this kind.

[1] The subject markers in this study had an undergraduate degree in the subject being marked or a cognate subject, but they did not have experience of teaching A-level and GCSE students.

Table 1 Categorisation of markers

	GCSE Mathematics	A-level Physics	Total
'Expert'	6 (3 female)	4 (1 female)	10 (4 female)
'Subject'	2 (1 female)	2 (0 female)	4 (1 female)
Total	8 (4 female)	6 (1 female)	14 (5 female)

All markers were briefed both on how to use the appropriate mark schemes accompanying the questions and on how to use the software package for marking from image. The software required that the markers marked their response samples 'item by item' rather than 'candidate by candidate'; that is, they marked one question answered by one candidate, followed by the same question answered by a different candidate, and so on until they had marked all responses to that particular question. After a short practice with the software and mark schemes, the reliability of all markers' marking was confirmed and each marker received individual feedback from a senior examiner.

As in the first study, the markers were then advised on how to provide concurrent 'think aloud' verbal protocols (Green, 1998) while marking in an experimental setting. After having marked for at least one day, the markers were asked to 'think aloud' whilst marking their selected response samples. Their verbal protocols were tape-recorded individually by one of two researchers and notes were made throughout the process. Occasionally, markers were asked for supplementary information. Subsequently, the tape recordings were transcribed.

Analysis and Findings

Initially, to obtain an overview of the data, the verbal protocol transcripts were given a broad read through by two researchers. This led to an overall judgement that the framework of five cognitive marking strategies identified in our initial study (Suto and Greatorex, *in press a, b*) could be applied meaningfully to the data. In a more detailed analysis of the transcripts, examples of the marking strategies were then identified.

Table 2 lists the cognitive marking strategies identified in each marker's verbal protocol. Extracts from verbal protocols which were interpreted as illustrating the presence of each of the cognitive strategies are given in Table 3.

Table 2: Cognitive marking strategies identified in each verbal protocol

Marker ID	Subject	Expert or subject marker?	Cognitive marking strategies used? (Y = yes; N = no)				
			Matching	Scanning	Evaluating	Scrutinising	No response
I	Maths	Expert	N	N	Y	Y	N
2	Maths	Expert	Y	Y	Y	N	Y
3	Maths	Expert	Y	N	Y	Y	Y
4	Maths	Expert	Y	N	Y	Y	Y
5	Maths	Expert	Y	Y	Y	Y	Y
6	Maths	Subject	Y	N	Y	Y	Y
7	Maths	Subject	Y	N	Y	Y	Y
8	Maths	Subject	Y	Y	Y	Y	Y
9	Physics	Expert	Y	N	Y	Y	N
10	Physics	Expert	Y	Y	Y	N	Y
11	Physics	Expert	Y	Y	Y	Y	N
12	Physics	Expert	Y	Y	Y	Y	Y
13	Physics	Subject	Y	Y	Y	Y	N
14	Physics	Subject	Y	Y	Y	N	N

As indicated in Tables 2 and 3, all five cognitive marking strategies were used in the marking of responses to both GCSE Mathematics and A-level Physics examination questions. This finding is in line with that of our initial study, where all five strategies were used for marking each of the subjects considered. (Although one ‘expert’ Mathematics marker participated in both of our studies, his verbal protocol data was not critical to this finding.) For Mathematics, all five strategies were used by both ‘expert’ and ‘subject’ markers. However, for Physics, while all five strategies were used by ‘expert’ markers’, the ‘no response’ strategy was not utilised by the two ‘subject’ markers.

Table 3: Extracts from verbal protocol transcripts exemplifying each of the five cognitive marking strategies

Strategy	Mathematics: 'expert' markers	Mathematics: 'subject' markers	Physics: 'expert' markers	Physics: 'subject' markers
Matching	*(5) 1.8 again*	*(7) And I look at the answer: it's incorrect.*	*(9) One right; one wrong.*	*(13) 10 to the -21, which is correct*
Scanning	*(5) I'm looking for now is $3x^3 + 15x$ somewhere on the page. There it is.*	*(7) [After looking at the answer line and finding the final answer incorrect] Looking for any working and there isn't any.*	*(10) There's nothing showing here except an H so that doesn't count as NR it just counts as 0.*	*(14) Just have a quick check that they've put the equation in but it's not going to affect anything. I just like to see that they've done it properly.*
Evaluating	*(4) 'To find the number of stars add the next two pattern numbers together.' Yeah, not quite right.*	*(6) I know that's no marks because that should be there and they would've got the mark for that one but that's pretty wrong.*	*(10) 'Charge on capacitor exponential decrease.' Yes that's OK so that's one mark.*	*(13) 'Greater time has elapsed from further objects as the light travels to earth.' So he's got greater time for travel which isn't bad. 'Distance is even greater' which again is the same point, but he hasn't mentioned the stretching of the wave length due to space expanding so that's just the one mark.*
Scrutinising	*(4) But I don't know how they've got it so I've just got to give it a bit of a think. '9, 3 then 1 each time. ' I really don't get why you've done that at all. Multiplied by 3, so I'm going to circle that bit and … does that pattern work? I've just got to check whether the pattern works every time so … I don't know where they've got that from so I'm going to give them 1 mark instead of the full 2 marks because there's something missing.*	*(6) It should be on, … It's meant to be um ….. a Y axis but I'm not sure, because they've put it on this side. It should be on that side mirroring that one there. I'm not sure about that one, I might refer that one, if that's OK? Yes. …. I don't know which one to grade for because they haven't really got one. That's a point, is it? I'm not sure about that I think.*	*(11) This one, they look as though they've use the right formula; I'll just see if they carried forward the error. No they can't have done, so they haven't actually used k in the formula at all in the end so that's 0.*	*(13) Um …. They've used the wrong value for 'g' here. We'll give … ah let's see…no he's …. yes, no he's given the actual answer he's…and also he's said it's approximately 180, so that's fine.*
No response	*(5) I can't even see whether he's put anything there, I don't think he has.*	*(7) Next one is blank.*	*(12) No response.*	

Note: The bracketed numbers preceding the transcript extracts are Marker IDs.

Discussion

This reanalysis of verbal protocol data from an earlier project had three aims. We investigated whether the five cognitive marking strategies identified in paper-based marking by expert GCSE examiners (Suto and Greatorex, *in press a, b*) were also used: (i) in A-level marking; (ii) by 'subject' markers; and (iii) when marking on screen. Our main finding was that, with the small exception of subject Physics markers not utilising the 'no response' strategy, all of the strategies were used in all three conditions. The single exception might well be attributable to the particular samples of candidates' responses marked by the two markers concerned.

There were several limitations to the study. First, since this was a reanalysis of data collected for other purposes, the selections of markers, examination questions, and response samples, were not made with the aims of this study in mind. The small sizes of the four groups of markers, and the lack of consistency among the response samples marked, prevented any meaningful quantification of strategy usages. Moreover, the lack of certain additional groups of markers (for example, A-level Mathematics and GCSE Physics markers; A-level Physics markers marking on paper rather than from image) prevented some properly controlled comparisons from being made. Secondly, there were several occasions during the qualitative analysis of the verbal protocol transcripts where it was unclear which candidate's response to which question was being considered by the marker. Unlike in our first study, only the question papers and transcripts (and not the candidates' responses) were available for use in our analysis. Together, the above limitations prevented us from linking the usage of each strategy or strategy combination with particular question types or response types, or with errors in marking. Similarly, we could not establish whether individual examiners favoured particular cognitive strategies.

Despite these limitations, our findings may prove useful. First, they validate the five marking strategies identified in our initial study (Suto and Greatorex, *in press a, b*), and indicate that these strategies were not specific to that single context. Having found them to be used in quite dissimilar marking settings, we suggest that the strategies are potentially quite general for GCSE and A-level shorter answer questions. However, it should be emphasised that our framework of strategies is not exhaustive. As explained earlier, Sanderson (2001) has proposed a marking model for marking A-level essays. At Cambridge Assessment, the Core Research Team has begun to conduct research exploring the differences in strategy usage among examiners marking both essays and shorter answer questions in an A-level subject. It is likely that this research will offer a closer exploration of the evaluating and scrutinising strategies. We hope that our colleague Vicki Crisp will be able to present this research at a

future IAEA conference. We anticipate that this research will continue to advance our knowledge of human judgement in marking.

Secondly, at a very general level, we found no evidence for striking differences in the cognitive marking strategies used by the 'subject' and 'expert' markers. It is worth noting that this finding is in contrast to those summarised by Weigle (1994) and others: in the marking of ESOL examinations, it was found that although there are some general differences between subject and expert markers in terms of the approaches used to assess writing, there are also some individual differences. All five of our strategies were used by both 'subject' and 'expert' markers, who apparently utilised both the quick and automatic 'System 1' thought processing, and the slower, rule-governed 'System 2' thought processing, posited in cognitive psychological theories of human judgement (Kahneman and Frederick, 2002). Our finding suggests that the potential for 'subject' markers to mark GCSE and A-level examination questions professionally should not be overlooked. Some research into what subject markers can mark has been undertaken by Royal-Dawson (2005) in Key Stage 3 English tests. However, further research is needed, and the validity and reliability of 'subject' and other categories of non-expert markers for GCSE or A-level would need to be confirmed. Within the Core Research Team, Irenka Suto and Rita Nadas are currently conducting research to explore which GCSE questions (if any) can be validly and reliably marked by subject markers. This research will investigate the role of teaching experience in marking different types of items in GCSE school examinations.

It is worth noting that although the subject markers in our second study appeared to use the same strategies as the expert markers, some strategies must entail the utilisation of slightly different combinations of information among each marker group. For example, subject markers do not have any teaching experience or past experience of marking other GCSE or A-level examination papers (which they made explicit), and therefore cannot make use of it when using the *Evaluating* strategy (Figure 4). Similarly, our finding of no striking differences in the cognitive strategies used to mark the GCSE Mathematics paper on screen and on paper should not be taken to mean that there are no differences *per se* between on screen and on paper marking. There are certainly some logistical and practical differences (O'Hara and Sellen, 1997; Greatorex, 2004), which have not been considered in our research.

Thirdly, some senior examiners in our first study suggested that the strategies should be made explicit in training courses for new examiners (Suto and Greatorex, *in press a, b*). Indeed, some senior examiners already advise examiners to use the strategies in an implicit manner (i.e. without using the term strategy or the terms matching etc). It has been

suggested that inexperienced examiners could have the opportunity to listen to the verbal protocol of senior examiners and to simultaneously see the associated script. This would give the new examiners insights into expert examining. Given that the same marking strategies appear to be used when marking on-screen, this idea could be elaborated upon and used in remote examiner training, whereby new examiners watch videos of a senior examiner's screen view as this senior examiner marks on screen (this would include the digital images of annotated scripts) whilst listening to the senior examiner's simultaneous verbal protocol. The utility of this idea would need to be tested.

Finally, an area of particular interest from a psychological perspective is the cognitive demand of marking short question responses on screen. Initially, examiners would need to become accustomed to e-marking software. It follows that although they would use the same cognitive marking strategies for on-screen and paper-based marking, on-screen marking could be more cognitively demanding than paper-based marking in its initial stages. Our cognitive marking strategies provide a benchmark against which the demands of future marking methods can be compared.

In conclusion, our second study has enabled us to understand further the similarities and differences between judgements made by expert and subject markers in paper-based and on-screen marking. We have outlined how our findings might impact upon examiner training and could influence how examiners are advised to make decisions, now and in the future.

Acknowledgements

We would like to thank Rita Nadas for preparing Figures 1 to 7. The original versions of the figures will be published in issue 2 of *Research Matters. A Cambridge Assessment Publication*. An adapted version of the figures has been reproduced here with the kind permission of, the Editor, Sylvia Green.

This research is based on examinations administered by Oxford, Cambridge and RSA examinations (OCR) and was funded by Cambridge Assessment (formerly, University of Cambridge Local Examinations Syndicate (UCLES)). The opinions expressed in this paper are those of the authors and are not to be taken as the opinions of Cambridge Assessment or OCR.

References

Baddeley, A. (1999) *Essentials of human memory* (Hove, Psychology Press).

Cumming, A. (1990) Expertise in evaluating second language compositions, *Language Testing*, 7, pp. 31-51. In Weigle, S. C. (1999) Investigating Rater/Prompt interactions in Writing Assessment: Quantitative and Qualitative Approaches. *Assessing Writing* 6(2) 145-178.

Greatorex, J. (2004) *Moderated e-portfolio project evaluation*, Evaluation and Validation Unit, University of Cambridge Local Examinations Syndicate. Available at www.ocr.org.uk

Green, A. (1998) Verbal Protocol Analysis in language testing research. A handbook. *Studies in Language Testing 5.* Cambridge: Cambridge University Press.

Huot, B. (1988) The validity of holistic scoring: a comparison of the talk-aloud protocols of expert and novice holistic raters, Unpublished PhD dissertation, Indianna University of Pennsylvania. In Weigle, S. C. (1999) Investigating Rater/Prompt interactions in Writing Assessment: Quantitative and Qualitative Approaches. *Assessing Writing* 6(2) 145-178.

Kahneman, D. & Frederick, S. (2002) Representativeness revisited: Attribute substitution in intuitive judgment, in: T. Gilovich, D. Griffin, & D. Kahneman (Eds) *Heuristics and biases: The psychology of intuitive judgment* (Cambridge, Cambridge University Press).

Laming, D. (1990) The reliability of a certain university examination compared with the precision of absolute judgments, *Quarterly Journal of Experimental Psychology*, 42A, pp. 239-54.

Laming, D. (2004) *Human judgment. The eye of the beholder* (London, Thomson).

Milanovic M., Saville, N. & Shuhong, S. (1996) A study of the decision-making behaviour of composition markers, in: M. Milanovic & N. Saville (Eds) *Studies in Language Testing 3: Performance testing, cognition and assessment - Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (Cambridge, Cambridge University Press/University of Cambridge Local Examinations Syndicate).

O'Hara, K. & Sellen, A. (1997) A comparison of reading paper and online documents. *Proceedings of the Conference on human factors in computing systems (CHI '97)*, p. 335–342. New York: Association for Computing Machinery.

Royal-Dawson, L. (2005) *Is teaching experience a necessity for markers of Key Stage 3 English?* QCA.

Sanderson, P. J. (2001), *Language and Differentiation in Examining at A Level*, Unpublished PhD dissertation. The University of Leeds, School of Psychology.

Shohamy, E., Gordon, C. and Kramer, R. (1992) The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, 76 (1), 27-33. In Weigle, S. C. (1999) Investigating Rater/Prompt interactions in Writing Assessment: Quantitative and Qualitative Approaches. *Assessing Writing* 6(2) 145-178.

Stanovich, K. & West, R. (2002) Individual differences in reasoning, in: T. Gilovich, D. Griffin & D. Kahneman (Eds) *Heuristics and biases: The psychology of intuitive judgment* (Cambridge, Cambridge University Press).

Suto, W.M.I. & Greatorex, J. (*in press, a*) What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process, *British Educational Research Journal*.

Suto, W.M.I. & Greatorex, J. (*in press, b*) A cognitive psychological exploration of the GCSE marking process, *Research Matters. A Cambridge Assessment publication*.

Suto, W.M.I. & Greatorex, J. (*in submission*) A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations.

Vaughan, C. (1992) Holistic assessment: what goes on in the rater's mind? In: L. Hamp-Lyons (Ed) *Assessing second language writing in academic contexts* (Norwood, NJ, Ablex).

Webster, F., Pepper, D. & Jenkins, A. (2000) Assessing the undergraduate dissertation, *Assessment and Evaluation in Higher Education*, 25, pp. 71-80.

Weigle, S.C. (1994). Effects of training on raters of ESL compositions: Quantitative and Qualitative approaches, PhD dissertation, University of California, Los Angeles. In Weigle,

S. C. (1999) Investigating Rater/Prompt interactions in Writing Assessment: Quantitative and Qualitative Approaches. *Assessing Writing* 6(2) 145-178.

Weigle, S. C. (1999) Investigating Rater/Prompt interactions in Writing Assessment: Quantitative and Qualitative Approaches. *Assessing Writing* 6(2) 145-178.

Yorke, M., Bridges, P. & Woolf, H. (2000) Mark distributions and marking practices in UK higher education, *Active Learning in Higher Education*, 1, pp. 7-27.

Dr Jackie Greatorex is *Principal Research Officer* of the *Core Research Team* in the *Research Division* of *Cambridge Assessment*. Jackie has been an educational researcher for 13 years. Jackie has published research papers about the comparability of grading standards and a novel method of writing grade descriptors based on empirical evidence for schools examinations. Additionally she has published research articles about the co-ordination/standardisation process and the reliability of assessment judgements in the contexts of school examination marking and vocational assessment.

Dr Irenka Suto is a *Senior Research Officer* in the *Core Research Team* in the *Research Division* of *Cambridge Assessment*. She joined Cambridge Assessment after conducting doctoral and post-doctoral research at the University of Cambridge, developing novel methods of assessing financial decision-making capacity and publishing many scholarly articles in this field. Irenka has contributed to several projects in her time at Cambridge Assessment, including research about examiners' cognition.