What does a good one look like? Marking A-level English scripts in relation to others

Victoria Elliott
Department of Education, University of Oxford, UK

Abstract

This article explores the use of *representativeness* as a guide to examining at English A-level through an analysis of two training days on two different modules. *Representativeness* is a cognitive heuristic which guides decision-making essentially by asking 'how much does this example look as if it belongs to this class of things?' A number of *representative* characteristics emerged during the training meetings including length, 'adult' writing and quality of written communication. The relation between *representativeness* and the mark scheme is also explored.

Keywords

examinations, marking, representativeness, comparative judgement

Introduction

This article reports data from a participant observation and discourse analysis study of two examiner training meetings in A-level English Language and Literature. It focuses on the ways in which examiners think about 'what a good one looks like', using, or challenging, a cognitive heuristic called '*representativeness*', which is essentially a question of how much an individual looks like an example of a wider category. This concept is used to consider some of the features of an examiner training meeting, and the ways in which examiners make judgements about English essays; it has the potential to bring together several different approaches to marking such as guild knowledge and construct-referenced marking. The purpose of this article is to explore the fruitfulness of this concept in thinking about the examining process.

The examiner training meeting is under-researched, partly because access to the training meeting is so difficult to obtain. In particular it is rare for someone

*Corresponding author: velda.elliott@education.ox.ac.uk

not associated with an Awarding Body to be able to research this aspect of assessment, and any insight into the process is valuable. In England, A-level English qualifications are set by four different Awarding Bodies, more commonly known as exam boards; this study took place in one particular Awarding Body.

The examining process is as follows: once students have taken examinations, the Principal Examiner or Examiners for a paper mark a selection of scripts to set the standard for that year, and to get an idea of the range of answers which are likely to occur. A few days after the examination, a training or 'standardisation' meeting is held for all the examiners who will be marking that particular paper; the Principal Examiner uses scripts from among those they have already marked to form the basis of their training. Examiners are introduced to the mark scheme and the standard, before going on to mark some further scripts at home to check that they have been 'standardised'. The training meeting is therefore a key part of ensuring parity between examiners, across schools and candidates. This process is common across A-levels and GCSEs in England and Wales, with a few exceptions where examiners undergo some online automated training before marking the scripts which judge if they are 'standardised'.

Training meetings follow a pattern of activity in which the Principal Examiner introduces examiners to some specific anchor scripts, talking through the decision-making process and the mark which was awarded. During the course of this process, the onus is gradually transferred on to the examiners. After this a set of 'training scripts' is worked through by examiners, still with pre-set marks awarded by the Principal and Senior Examiners, but also with discussion based around the trainee examiners' thought processes.

It is worth noting that, although the meetings described here are from before the recent reforms of A-level, the processes will carry over into the new specifications. The nature of examination answers (essays) remains the same, and the assessment criteria are closely related to the previous ones. The conclusions reached here are not intended to be normative: this is not a training manual for examiners, but is instead an exploration of what examiners talk about and draw on when making judgements of examination essays.

The role of the mark scheme

The mark scheme holds a central position in the marking process, and is designed to hold the absolute standard against which all marking judgements are made. In essay subjects such as English, the mark scheme is composed of generic descriptors of levels of attainment and a passage of question-specific 'indicative content'. (Although not all boards or specifications include this, the units in this study did.) The generic descriptors remain the same from year to year; teachers and candidates are well aware of them. Indicative content, used widely in the host organisation's specifications, is neither exhaustive nor a list of information required for full marks. It is simply an acknowledgement that examiners may be marking examinations on topics within their subjects with

which they are not familiar. The generic descriptors are divided into 'bands' of marks representing different levels of achievement; examiners do not know how many marks equate to what final grade, since the setting of grade boundaries is done at a later date, more or less entirely in accordance with statistical guidance. A-level examinations aim to assess several distinct 'Assessment Objectives', which can be weighted in different combinations depending on the particular examination (within certain limits set for the qualification as a whole). This adds to the level of complexity which is part of examining, as markers may not only have to make judgements against several different sets of criteria but also may have to make judgements against one set of criteria which involves two or more differently weighted Assessment Objectives.

In the assumed model of examination marking, the mark scheme is considered side-by-side with each essay, and drives the process of judgement. The stated purpose of the standardisation meeting is to ensure that examiners have a 'common well-founded understanding of the mark scheme', yet in my data there is a lack of explicit focus on the mark scheme as a document, or on understanding it *per se*, as opposed to seeing how it has been operationalised by senior markers. The mark scheme is generic and it is abstract; much of the activity of the training meeting is devoted to seeing how it can be made concrete.

The mark scheme represents another level of difficulty in the marking process in its use of modifiers to indicate levels of skill, so that examiners must know the distinction between, for example, 'mainly accurate', 'generally accurate' and 'accurate'. This is something which can only be done by exemplification, or by understanding an underlying construct; there is no way of explaining verbally what the distinction would be. This is one of the features that points to the existence of 'guild knowledge' (see below).

Standardising meetings and marking English

There are very few studies which deal with the standardisation process in examining, and in particular the training meeting, which has remained a closed process. Studies which deal with the technical issues of examination of English are very few and far between, and for that reason much of the literature reviewed here draws on other subjects. There are two detailed studies which have looked at standardising meetings: Daugherty (1988) and Sanderson (2001). Daugherty (1988) sought to establish the process by which candidates were awarded grades in Geography O level examinations in 1984 and 1985, and drew on observations of the question-setting meetings, the standardisation meetings and the grade awarding meeting (at which grade boundaries are decided). Despite the time which has passed since this research and its context in Geography it is worth considering, since it is the most complete account available. Sanderson (2001) draws on eleven years of 'participant observation' as a Sociology examiner at GCSE and A-level, as well as the observation of the pre-standardising and standardising meetings for one examination each of Law

and Sociology for the main data gathering phase, but is a descriptive outline of the training meetings rather than a detailed analysis.

Daugherty established three key features of the 'examiners' conference' (which is referred to in this study as the standardisation or training meeting): the discussion is focused on actual scripts; all possibilities are considered; and discussion continues until there is an agreed marking scheme (1988:17). The examiners worked on a draft marking scheme, with scripts of actual answers, and ensured that all possible answers to a question had been considered; in one of the two meetings the aim was for an exhaustive list, while in the other meeting, examiners were encouraged to use their own professional judgement. Sanderson (2001) also reports the alteration of mark schemes during pre-standardisation meetings (in either law or sociology; he did not distinguish between the two in his data). This differs from the current practice in English (seen in the data, and through discussion with a wide range of examiners), in which the mark scheme is set before the examiners meet, and there is no consideration of other possible correct answers; this is due to the difference in subject (as confirmed by current geography examiners), and the nature of short answer questions, rather than to changing practice over time.

Daugherty considers the chief examiners to have 'retained the high ground of the decision-making process' (1988:24) although they were challenged at times by the other examiners, and he describes the style of the decision-making in one meeting as 'consensus'. On a small number of occasions, the group of examiners settled on a decision by a majority vote. The process which Daugherty describes is one in which there are many inputs to the overall process of judgement; 'each participant in the examining process brings to it a personal perception, based on experience, which will influence their judgement' (1988:46).

In his data, Sanderson notes variation in how much challenge was allowed to the Principal Examiner's view, but adds the caveat that in subjects where there are very large numbers of candidates, and therefore examiners, 'transmission of judgement tends to be rigidly hierarchical' (2001:103). He cites an interview with a Senior Examiner for English in support. Although neither of the units considered in this paper was very large, the subject mentality remained.

English itself presents a special case for marking, because of the subjective and somewhat affective element involved in the judgements to be made (Elliott, 2013, 2014). Britton's work in the 1950s and 1960s effectively demonstrated the difficulty of getting markers to agree on the quality of English essays (in his case, 'compositions'). Britton's (1964) evidence suggested that 'impression' marking by multiple markers, averaged out, would be a more reliable and valid method of marking English than via the analytical mark schemes used by examination boards. In part an imbalance was created by the separating out of features to award discrete marks. Gibbons and Marshall (2010), reporting a

collaborative assessment study with English teachers in London, showed that even with a detailed analytical mark scheme created by teachers themselves, a range of marks was still given to the same texts, suggesting difficulties with creating a common understanding of certain qualities or characteristics, or with the application of weighting to different features. As Marshall argued later, 'English teachers have a tendency to want to consider the whole of a piece rather than look at its constituent parts' (2011, p. 12).

One explanation which has been drawn on to consider how teachers come to a common understanding over a period of time is 'guild knowledge', as coined by Sadler (1989) and considered more recently by Marshall (2011). Here the indoctrination of the teacher into the 'guild' gives them access to a shared knowledge; Sadler discusses it in the context of formative assessment which, by:

> *providing guided but direct and authentic evaluative experience for students enables them to develop their evaluative knowledge, thereby bringing them within the guild of people who are able to determine quality using multiple criteria*
>
> Sadler 1989: 135; italics in the original.

In the same way, teachers gradually gain guild knowledge by shared assessment and discussion, and through a set of shared values of what it means to be 'good' at the subject. (It is this guild into which assessment for learning seeks to bring the student.) Marshall, while adding the critique that this does not account for potential differences in points of view about a subject (2011), links this to Britton's (1964) 'impressionistic' marking, in that both involve a shared knowledge which it is not necessarily possible for the teacher or guild member to articulate (Polyani's 'tacit knowledge', 1958). Wiliam, drawing on an alternative terminology but still echoing these concepts, argues that teacher assessments of this kind 'are construct-referenced assessments, validated by the extent to which the community of practice agrees that the student's work has reached a particular implicit standard' (1998:7). The concept of the construct-referenced, rather than criteria-referenced, assessment, is one in which scripts are held up against a mental construct of 'what an A is'.

Representativeness, mental frameworks and comparison

The original study from which this data is drawn sought to explore the decision-making abilities of A-level examiners in the light of several potential frameworks taken from laboratory studies in cognitive psychology, to see if they were applicable in a real world context. This led to the concept of representativeness. Kahneman and Tversky's theory of heuristics and biases provides an explanation of the ability of the human to make intuitive judgements when the cognitive load is beyond their rational capability (as it might be when being asked to consider a mark scheme, a candidate essay and a range

of anchor essays, under time pressure), using three heuristics which involve assessing the information that is available (*availability)*, judging its *representativeness* against samples from their prior experience (*representativeness)* and then adjusting to fit the scenario currently in question (*anchoring and adjustment)*. Gigerenzer and Gaissmaier define heuristics as 'efficient cognitive processes, conscious or unconscious, that ignore part of the information' (2011: 451). The 'biases' are where error creeps in, so that there is a trade off between speed and accuracy, a dilemma that English teachers know well. Biases often occur when cues for judgement are wrongly weighted in the judgement process (Kahnemann & Frederick, 2002). However, heuristics are a valid, well-used, automatic strategy in human judgement processes.

Representativeness is a heuristic in which items which appear similar are assumed to have the same characteristics, and an item which appears to fit into a group is assumed to have the characteristics of that group. This results in a situation in which 'some probability judgements (the likelihood that *X* is a *Y*) are mediated by assessments of resemblance (the degree to which *X* "looks like" a *Y*)' (Kahneman & Frederick, 2002: 49–50). This may seem a reasonable approach to take in marking, in which we might ask 'to what extent does this look like an A grade?', but the problem lies in the fact that 'look like' is not necessarily based on the key characteristics of the category. The example given by Tversky and Kahneman is as follows:

> *Steve is very shy and withdrawn, invariably helpful, but with little interest in people, or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.*

Is Steve a farmer, salesman, airline pilot, librarian, or physician? The *representativeness* heuristic means that 'the probability that Steve is a librarian, for example, is assessed by the degree to which he is *representative* of, or similar to, the stereotype of a librarian' (Tversky and Kahneman, 1974: 1124). This is analogous to the concept of illusory covariance or the 'halo effect', which was the name Thorndike (1920) gave to the tendency for an overall impression to affect our judgement of specific characteristics, so (for example) more attractive people are also judged to be more competent. Similarly in examining, a script whose content is worthy of an A grade may be under-judged because the handwriting and spelling are sub-standard. Equally a well-presented script may appear *representative* of a higher grade than its contents deserve. There have been some studies which have shown that heuristics can be more accurate rather than less, and there comes a point where too much information reduces the accuracy of judgements (Gigerenzer, 2015), which nicely echoes the findings of Britton, discussed above.

Researchers in assessment, rather than cognitive psychology, have considered the existence of a mental framework into which examiners fit each script they

read, according to its quality. Vaughan, for example, in her paper on rating in EFL, suggests that when papers are read quickly, one after another, as they are in a holistic assessment session, they become, in the rater's mind, one long discourse (1991: 121). She cites the informal comparative statements made by seven of the nine raters as evidence of this. It also resonates with the idea of an overall mental frame of reference into which essays are fitted. The concept of internalised, tacit standards has been accepted by Crisp (2010a) drawing on the psychological terms of Rosch (1978), to describe the 'mental models of likely typical responses' as 'prototypes' (Crisp, 2010a:21).

Wiliam (1998) termed the use of such mental frameworks 'construct-referenced' assessment; i.e., assessment done with reference to the 'construct' or personal understanding of what it means to be a given grade. Marshall (2001), intending to test this, used a qualitative methodology, observing several in-school meetings at which groups of teachers standardised their coursework marking by grading papers supplied by the exam boards. She found that teachers preferred to think in terms of grades rather than marks and quotes a rather telling remark, from one teacher 'I instinctively know what it is and adjusted the marks accordingly. This screams D' (Marshall, 2001: 53), a comment which exemplifies the use of constructs of grades by teachers. These constructs were used even though they clashed to some extent with the specific criteria laid down in the mark scheme, with the teachers refusing to be 'bogged down' (in their words) by the criteria. Some graders interviewed by Baird and Scharaschkin (2002) suggested that their decisions of grade-worthiness were based on a gut feeling or instinct (five out of ten business studies examiners but only two out of nine English). The majority also suggested that their teaching experience helped them to form an opinion as to what a B-grade required;. Baird, asking examiners to judge whether scripts were worthy of an E or not, with various different exemplars, concluded that English literature examiners were 'probably using an internalised notion of standards to carry out a grading categorisation task' although these internalised standards were somewhat fuzzy and modifiable by the application of different reference material (2000:98).

However, the reference to mental constructs or frameworks is not always held to be safe; Pollitt uses the word 'imagined' to describe them and asks 'what is the imagined performance that properly embodies a particular verbal descriptor?' (2004:6); it is the nature of the mental framework that it remains unexplicated and internal, so that there is no way of standardising the framework. *Representativeness* is not superior to these concepts, but may be useful to illuminate the practices of examiners, and to draw together some of the disparate conceptualisations which have been discussed here, and may help to explain the mechanism by which 'mental frameworks' work.

In his influential book *Human Judgment: The Eye of the Beholder*, the psychologist Donald Laming stated unequivocally: 'there is no absolute judgment. All

judgments are comparisons of one thing with another' (2004:9), which fits with the relativism of the criteria in the mark scheme. An emphasis on relativity can be confirmed by studies such as that of Crisp (2010a) which compared 'think aloud' protocols of groups of examiners making grading decisions and the same examiners individually marking; in both activities, comparison, either to another script or to another response within the same script, occurred regularly, at a frequency of 0.66 and 0.69 instances per script for marking and grading respectively. Crisp (2010b) includes comparison among the cognitive processes of markers without question. Comparison is the basis of the *representativeness* heuristic; judgement is based on the similarity of the specific example compared to the larger group.

Gill & Bramley (2008) reported a study which used history and physics A-level scripts to investigate absolute and relative judgements of examiners, asking them to make three judgements: an absolute judgement of the grade a script was worth; a relative judgement of which of two scripts were better in terms of quality; and an assessment of their confidence in each judgement. They found that examiners had difficulty in accurately judging the absolute grade-worthiness of scripts, although the history examiners (a subject more similar to English) were more able to do so than physics. The relative judgements were, however, more accurate than the absolute judgements. Both sets of examiners had more confidence in their relative than their absolute judgements. These findings seem to support Laming's view.

The data reported in this study, therefore, considers the role of *representativeness* in how examiners make judgements about scripts, and the ways in which they relate to the mark scheme.

Method

Two day-long examiner training meetings for A-level English Language and Literature were observed, and recorded. One meeting was an online meeting, in which the whole day was digitally recorded. The other was a small meeting of examiners of a new module, examined for the first time, conducted in person. I was present throughout both meetings and had full access to all materials used by the examiners, including training scripts and mark schemes. Access was obtained to the meetings through gatekeepers at the top of the organisation, undoubtedly helped by my university association, and I was allowed free access to the meetings, with no restrictions placed on me beyond preserving the anonymity of the organisation and the individual examiners. In the 'in-person' meeting I was much more obviously present, and more of a participant observer. In the online meeting my presence was known to all participants, but I didn't speak. One examiner from the live meeting also completed a think aloud study during live marking (an approach successfully used to study examining, e.g. Suto & Greatorex, 2008). In total, fifteen hours of data make up this study, which reflects the difficulty of access; some data is always better than none, however.

The dominant methodology used in studies of examiner thinking is Verbal Protocol Analysis (VPA) which is also known as the 'think aloud' method. VPA is based on the assumption that 'an individual's verbalisations may be seen to be an accurate record of information that is (or has been) attended to as a particular task is (or has been) carried out' (Green, 1998:1-2). Graesser *et al.* approved the use of 'think aloud' protocols as providing 'a very rich source of data for discovering possible comprehension strategies and for testing detailed claims about the representations that enter the reader's consciousness' while warning that 'protocols do not reliably tap unconscious comprehension processes' (1997:166). Both examiners and their trainers during the meeting seek to make explicit their thinking, the former as a means of evidencing their competence, and the latter as a model. This makes their talk a suitable data set for VPA.

Data were transcribed and analysed using codes drawn from the literature; these included looking for evidence of cognitive heuristics such as representativeness, for use of construct referenced assessment, and for holistic versus analytical (broken down) judgements. A stage of inductive coding was also carried out, which allowed other aspects of the data to be elicited which were not drawn from the theoretical framework; this raised, for example, the issue of comparison and the reference points used during relative judgement. The aim of this approach was not only to enable the exploration of theoretical concepts within a 'real world' situation, but also to ensure that they did not obscure anything else that might arise out of the data.

Setting the scene

English 1

A small team of approximately twelve examiners together with a Principal Examiner took part in this day-long meeting. (I was not the only 'observer' in this meeting; there were two or three employees of the Awarding Body, and no distinction was made on the online portal between examiners and observers. Twelve of the attendees participated in the discussion without being identified as exam board employees, and were therefore regarded as examiners.) Although it followed the same generic format of briefing from the exam board, then training scripts followed by standardising scripts, the distinction between the two types of scripts was blurred, so that little formal standardising took place. There was considerable discussion of the features of each script and the mark which it should be assigned, chaired and managed by the Principal Examiner. She invited participants to talk, or gave the floor to individuals who showed an electronic 'hands up'. The marks which had previously been awarded by the Principal Examiner were, however, not negotiable.

There was a sense of fraughtness over the technology; it was the first time this paper had had an online standardising meeting. A considerable delay in beginning the training aspect of the meeting was caused by various team members,

including the Principal Examiner, being unable to use the system or to access some pdf files. Once they had begun, the Principal Examiner was anxious to keep going without breaks so that further technological difficulties did not occur during a pause; the meeting ran smoothly from then on. The paper consisted of a few short answer questions followed by an essay. The short answer questions do not form part of the discussion here.

English 2

The smaller, physical meeting was held in a conference room in an office building in London. The room, the size of an average secondary classroom, was arranged with groups of tables pushed together; all participants in the meeting sat around three double tables pushed into one large one. The meeting was opened by a Standardising Officer from the exam board, who registered participants and ensured that they had travel expenses forms. Once his briefing was delivered he left the room, although he remained available for support.

Actually taking part in the meeting were the Principal Examiner and his team of two examiners, one of whom was an experienced marker who acted as a Team Leader for another module. In addition, the Chair of Examiners for the subject was present (who oversees the Principal Examiners for each specification within that subject), and a Senior Examiner who had been involved in writing the specification, question paper and mark-scheme, but was not marking this time. (She was also a Chief Examiner for another English syllabus). The five participants all knew each other well and had a professional relationship stretching back some time.

The Principal Examiner did not follow an authoritarian regime for this meeting. Discussion was egalitarian in nature, and there the standard was not unalterably set; the Principal Examiner was open to changing his views. Although he chaired the meeting, there was no sense that examiners were being tested on their ability to reproduce his marking standard, but rather that they should come to a consensus on an agreed standard. This meeting approached a 'consensus' or 'contest' model of decision-making; the tone was for the most part, 'sweet reasonableness' (Christie & Forrest, 1981:35) but participants were unafraid to challenge each other's perceptions. The last word was had by the Principal Examiner who chaired the meeting; he took great account, however, of other participants' opinions.

Representativeness

The phrases 'looks like' or 'feels like' are characteristic of the application of the *representativeness* heuristic, and it was particularly in evidence in relation to top scripts. Other phrases also suggested that examiners characterised the 'feel' of an essay in a variety of ways which enabled them to categorise the scripts as an aid to putting them in the correct bands. The examiners in the standardising meeting for English 2 were particularly concerned with the 'checklist

approach' towards writing the essays, a characteristic they saw as *representative* of lower levels of achievement.

The length of quotations was clearly *representative* for the examiners; the think aloud examiner frequently remarked on the presence of over-long quotations without accompanying explication by the candidate during her recordings, and during the English 2 meeting she commented 'I was a little uneasy about the length of the quotations.' Quotations were an easy source of *representative* characteristics; the examiners also regarded the presence of 'nicely embedded' quotations as indicative of a higher order of skill.

Length was clearly considered to be a *representative* characteristic; in particular shortness is *representative* of lower grades. The following exchange in the English 2 meeting exemplifies this:

Principal Examiner	*Can we look at J? I found this a bit of a problem one – it's a very short answer.*
Examiner	*Oh golly.*
	[*they read*]
Examiner	*Good stuff though. Hard to take exception to any of it. Hardly a word wasted. I can't believe that wasn't planned. They probably spent more time thinking about it than writing.*
Principal Examiner	*But it's short, isn't it?*
Examiner*	*It is, but it's small writing.*
Examiner	*It's literally getting twice the number of words to the line.*
Senior Examiner	*It's discriminating, no doubt, not wasting a word.*

Although the Principal Examiner was focused on the shortness of the answer, the other examiners defended it not only on the basis of the content but also on the grounds that it was not as short as it appeared, because of the handwriting. They continued to delineate the essay's positive points after this exchange, until at the decision the Principal Examiner admitted 'I was thrown by the apparent brevity' – which is in accord with Crisp's (2010b) finding that when the length of a response does not correspond to an examiner's expectations, that becomes a factor in their decision-making. The team decided this essay was 'deceptive' and that it should be put away for the summer standardising meeting as a good example script. Despite its apparent length, the essay was placed in the top band. It was clear that the Principal Examiner had intended the essay to be a sample of a poor essay, and his emphasis on the length indicates that this was the characteristic which had led him to this

conclusion; it is also clear that the principle that shortness is indicative of low quality was not overturned here, as the examiners established that the essay was longer than it looked.

It is interesting to note ithat *representative* characteristics were not usually attached to a specific band of marks, but rather suggested a higher, lower or middle ranking essay. The only characteristic which was repeatedly connected to a specific band was the word 'adult', as in 'this is written by an adult' (metaphorically) or 'the writing is completely adult'. This was a term of approval linked to essays in the top-band, and particularly those given full marks. In general, however, *representative* characteristics did not appear to be tied to levelling decisions, which made them both of less use to examiners and less likely to lead to bias. Having made a judgement on representative grounds, examiners had to resort to the mark scheme to fine tune their decision (reflecting *anchoring and adjustment*, another heuristic, at work). This may be related to the fact that although there is intended to be a qualitative difference between bands, examiners in fact saw it in terms of gradations rather than categories, no doubt encouraged by the use of slightly differing adverbs in the bands of the mark scheme. No one *representative* characteristic seemed to drive any single decision, although one was particularly influential in adjusting the decision: Quality of Written Communication (QWC).

Quality of Written Communication

Quality of the writing (QWC) was an aspect of the scriptsfrequently noted by examiners on English 1 in their discussion. For English 2 QWC was formally assessed, as one of the Assessment Objectives included the requirement for 'accurate, coherent written expression.' It was an Assessment Objective (AO) with a low weighting, responsible for a very small proportion of the marks available for that paper. The examiners on both papers, however, used QWC as an indicator of the quality of the essay.

The examiner providing think aloud data frequently commented on the 'style' of the essays she marked; however, in her decision-making for the AO, which mentions 'written expression', she focused exclusively on the other elements in the AO description, particularly the range of literary and linguistic terminology which had been used. The 'style' comments formed instead a general impression of the quality of the essay early on; for an English examiner this may be closely related to, and aligned with, the skills which are actually supposed to form the basis of the assessment.

Essays which were excellent, and worthy of full, or almost full marks, however, were often identified first by their writing style (related to the discussion of 'adult' above), as an easy marker. 'Fluid' and 'sophisticated' styles were both seen as *representative* of the top level. The best practitioners of QWC were

also likely to have structured their essay, argument and content to the best effect, so that it became a proxy for the other skills which were formally examined on the mark scheme. Examiners were warned against candidates who *appeared* to be demonstrating QWC, through their use of signposting using discourse markers, for example, but who were not; so to be careful to establish when QWC is commensurate with the other skills and when not.

For the standardising meeting on English 1, QWC was an integral part of the commentary delivered by the Principal Examiner to explain her decision, particularly where it might be deceptive ('a bit wordy, but it's talking about subject specifics'), which suggested an acknowledgement on the part of the Principal Examiner that these characteristics might be used to make judgments.

Quality of Written Communication is a characteristic of scripts which is relatively easy to see and judge, as an automatic process, for English examiners. It seems likely that for some it becomes a *representative* characteristic, and as such principal examiners feel the need to warn against its potential for causing bias if wrongly applied. The data did not support Crisp's (2010b) finding that language became a more manifest criterion when there was some difficulty with the way in which a candidate had expressed themselves, with corresponding problems with the interpretation of the answer by the examiner: language is always a manifest criterion for English subjects.

Imagined representative characteristics?

Not all representative characteristics seemed to come from real scripts; where they felt the need, examiners would compare scripts with an imagined other. Alistair Pollitt (2010) alleges that in the absence of physical scripts to compare, examiners will resort to ideas of what a given level 'looks like'. It is also suggested by the 'prototypes' of Crisp's (2010a) model, the 'mental models of likely typical responses'. There is certainly much comparison throughout the data with hypothetical scripts and the change in marks that would result if a script looked a little different; examiners also constantly refer to what 'a better candidate' would do, or what an ideal answer would look like.

Examiners often compared a script at hand with an idealised one, so an examiner can comment 'better scripts would have said that it wasn't archaic at the time' (Senior Examiner, English 2). The anticipated shape of the field of answers is also considered, again showing that each answer is considered relative to the entire field; the question 'are we going to see ones which do have more on the language?' is answered with 'I don't think we will' (Senior Examiner and Principal Examiner, English 2); although the Principal Examiner has a slightly greater idea of what the field looks like, it is still not enough to make this much more than speculation. This kind of speculation is important in establishing what is *representative*; the imagined characteristics of a top-band essay have the potential to bias judgement of essays which deserve the top marks but do not look like that imagined best.

Talking schematically

Despite the apparent use of the *representativeness* heuristic, all examiners, consciously or not, echoed the terms of the mark scheme in their talk. They lifted phrases or produced paraphrases that mimicked it closely, when discussing papers. When marking was carried out on physical papers, examiners were told to refer to the mark scheme in their annotations, which justify their decisions to supervisors and to schools; papers are more frequently scanned and marked on screen, often without annotation, now, as was the case in these two examinations.

In English 2, the vocabulary of all examiners frequently referenced the mark scheme in their discussion of scripts, often weighting one aspect against another ('they know a lot of terms but it doesn't actually analyse' Examiner, English 2), but their discussion was wide-ranging and it is unsurprising that many of the terms which they used also appeared on the mark scheme, given that they are drawn from the common lexis of the subject. 'Analysis', for example, is a major concern of English and appears on a majority of A-level mark schemes. Similarly 'context' was mentioned even when the Assessment Objective under discussion did not assess it. This may suggest that these teachers were marking holistically, but terms which were not such staples of English teaching were not raised so much; 'attitudes and values', a key term for the mark scheme for English 2, was mentioned by name on only two occasions during the meeting.

The nuances of the adverbs used in the mark scheme and their prominence were not necessarily alien to the examiners, as one, again in English 2, commented that a script was 'probably more implicitly than explicitly analytical but it is "consistently"'. Later the Principal Examiner tells the team 'I'm still looking for that "precisely analytical"', and none of them queries the shades of meaning in these terms. Sometimes, however, they do cause difficulty: the Principal Examiner of English 2 on script describes himself as 'sort of hovering between 8 and 9 [marks] because it says "significant range of literary and linguistic" [terminology].'

The think-aloud examiner marked with the mark scheme as a constant reference. At the beginning she took phrases from the mark scheme and assessed the essay according to the presence or absence of each, as if working through a checklist of 'context', 'critical understanding', 'terminology' *et cetera*. She engaged repeatedly with the meaning of the terms in the mark scheme, asking herself questions: 'is it integrated?' She spent a great deal of time in her first 'think aloud' session trying to identify if certain aspects of the script equated to the terms on the mark scheme.

> *Is there a range of relevant contextual – very little [almost half a minute of silent thought] doesn't even use the word Victorian or – there was a mention of America… seen by society, but it doesn't*

> *actually say what society, there's no indication of time or f- there's just nothing there on context. Oh, except there was that ref— reference to performance which is a form of context isn't it…. "some awareness of context", yeah well that is some awareness isn't it?*

This shows a range of ideas about what constitutes 'context', running through them to see if the essay could satisfy the requirement. The examiner eventually identified something which satisfied her as context, and measured it against the wording of the mark scheme.

Conclusions

This aticle has explored the training of examiners and their marking of A-level English essays through the use of a particular conceptual framework. Using such a framework as a lens to examine data has the potential to elicit new and interesting findings, but the limitation is that it can also constrain the range of knowledge which results. *Representativeness* seems to be a fruitful concept in relation to this data, but it is certainly not the only concept which could be of interest. The cognitive process of examining is a complex one and there are many other elements which could be explored; the limitations of a single article limit exploration of, for example, the other heuristics, or the use of comparison by examiners, both of which interact with representativeness. This is also just one of the potential sources of data: the examiners' reports which provide a summary of the issues raised after each time an examination is set will also be of interest to teachers looking to understand the A-level assessment process, particularly in reference to the content which students' essays should contain.

The *representativeness* which an A-level English script displays of a mark category (a band) can be an important part of how examiners respond to it. An important part of the training meeting seems to be identifying what are safe and unsafe signals to follow in terms of the characteristics which are supposed to be *representative* of a given level. *Representativeness* does not seem from the observed behaviour of examiners to be something which they follow automatically, and indeed it does not correspond to specific marks or bands of marks, but instead to something more akin to a rough rank ordering. This falls within an acknowledged technique of examiner training, namely situating scripts within a framework of anchors, with which examiners can later compare live scripts. Since examiners will also compare with imagined scripts, it appears that they do have 'prototypes' in mind when marking, and that they do use a mental framework of construct-referenced assessment to some extent, although that construct is reinforced and supported by the mark scheme and the previously marked scripts with which examiners are supplied. The mark scheme provides a way of fine-tuning decisions which are made by 'feel', explained in cognitive psychology as being attributable to cognitive heuristics; this 'feel' is often seen in the literature around English teachers as being 'guild knowledge' or an internalised idea of the standard.

While the units discussed here were both from specifications in the integrated subject English Language and Literature, the essay format and the general key skills for English are also applicable in English Literature and English Language specifications. Where questions are more knowledge-based (as they might be in some Language essays) there are likely to be differences in the approaches that markers take. Raikes, Fidler and Gill, for example, note the relative ease of marking highly structured factual essays with a prescriptive mark scheme (in Psychology) against contexts with 'less constrained essays and marking' (2010: 26). However, in the context of literature essays, whether with a more linguistic focus or not, there is a definite sense of some key characteristics which examiners consider to be *representative* of 'what a good one looks like': 'adult' or sophisticated writing; length; and quality of written communication. It seems likely that such attributes will survive the reform to the new A-level s, standing beside, rather than above, the new mark schemes.

Acknowledgements
I am grateful for the ESRC doctoral studentship which funded the collection of this data. I would also like to thank the examiners who gave me their time and agreed to participate in the research.

References

Baird, J.-A., (2000) 'Are examination standards all in the head? Experiments with examiners' judgements of standards in A-level examinations', *Research in Education*, 64, 91–100

Baird, J.-A. & Scharaschkin, A. (2002) 'Is the Whole Worth More than the Sum of the Parts? Studies of Examiners' Grading of Individual Papers and Candidates' Whole A-Level Examination Performances', *Educational Studies*, 28(2), 143–62

Britton, J. (1964) *The multiple marking of compositions*. London: HMSO

Christie, T., & Forrest, G.M. (1981) *Defining Public Examination Standards*. Schools Council Research Studies, London: Macmillan Education

Crisp, V. (2010a) 'Judging the grade: exploring the judgement processes involved in examination grading decisions', *Evaluation & Research in Education*, 23(1), 19–35

Crisp, V. (2010b) 'Towards a model of the judgement processes involved in examination marking', *Oxford Review of Education*, 36(1), 1–21

Daugherty, R. (1988) *Examining Geography at 16+ A study of decision-making in two geography examinations*. London: Secondary Examinations Council

Elliott, V. (2013) "Empathetic projections and affect reactions in examiners of 'A' level English and History", *Assessment in Education: Principles, Policy & Practice*. 20(3), 266–280. DOI: https://doi.org/10.1080/0969594X.2013.768597

Elliott, V. (2014) "MARC: A Thought Experiment in the Morality of Automated Marking of English",*Changing English: Studies in Culture and Education*, 21 (4), 393–401. DOI: https://doi.org/10.1080/1358684X.2014.969000

Gibbons, S. & Marshall, B. (2010) 'Assessing English: a trial collaborative standardised marking project', *English Teaching: Practice and Critique*, 9(3), 26–39

Gigerenzer, G. (2015) *Simply rational: Decision making in the real world*. New York: Oxford University Press

Gigerenzer, G., & Gaissmaier, W. (2011) 'Heuristic decision making', *Annual Review of Psychology*, 62, 451–482

Gill, T., & Bramley, T. (2008) *How accurate are examiners' judgements of script quality? An investigation of absolute and relative judgement in two units, one with a wide and one with a narrow 'zone of uncertainty'*. Paper presented at the British Educational Research Association annual conference, Edinburgh, September 2008

Graesser, A.C., Mills, K.K. & Zwaan, R.A. (1997) Discourse comprehension, *The Annual Review of Psychology*, 48, 163–189

Green, A. (1998) *Verbal protocol analysis in language testing research a handbook*. Cambridge: CUP

Kahneman, D. & Frederick, S. (2002) '*Representativeness* revisited: attribute substitution in intuitive judgement', in T. Gilovich, D. Griffin, & D. Kahneman (eds). *Heuristics and Biases: the Psychology of Intuitive Judgement*. Cambridge: CUP. 49–81

Laming, D. (2004) *Human judgment: the eye of the beholder*. London: Thompson

Marshall, B. (2001) 'Marking the essay: teachers' subject philosophies as related to their assessment', *English in Education*, 35(3), 42–57

Marshall, B. (2011) *Testing English*. London: Continuum

Pollitt, A. (2004) *Let's stop marking exams*, paper presented at the IAEA Conference, Philadelphia

Pollitt, A. (2010) *How to assess writing reliably and validly*, paper presented at AEA-Europe, Oslo

Polyani, M. (1958) *Personal Knowledge: towards a post-critical philosophy*. Chicago: University of Chicago Press

Raikes, N., Fidler, J. & Gill, T. (2010) 'Must examiners meet in order to standardise their marking? An experiment with new and experienced examiners of GCE AS Psychology, *Research Matters*, 10, pp. 21–7

Rosch, E. (1978) 'Principles of Categorization', in E. Rosch & B.B. Lloyd, (eds) *Cognition and Categorization*. Hillsdale: Lawrence Erlbaum Associates, 27–48

Sadler, D.R. (1989) 'Formative assessment and the design of instructional systems', *Instructional Science*, 18, 119–144

Sanderson, P. (2001) *Language and Differentiation in examining at A-level*. Unpublished PhD thesis, University of Leeds

Suto, W.M.I., & Greatorex, J. (2008) 'What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process', *British Educational Research Journal*, 34(2), 213–233

Thorndike, E.L. (1920) 'A constant error in psychological ratings', *Journal of Applied Psychology*, 4, 469–477

Tversky, A. & Kahneman, D. (1974) 'Judgement under uncertainty: heuristics and biases', *Science*, 185, 1124–1131

Vaughan, C. (1991) '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 Publishing Corps

Wiliam, D. (1998) 'The validity of teachers' assessments', paper presented at the 22nd annual conference of the International Group for the Psychology of Mathematics Education, Stellenbosch, South Africa

