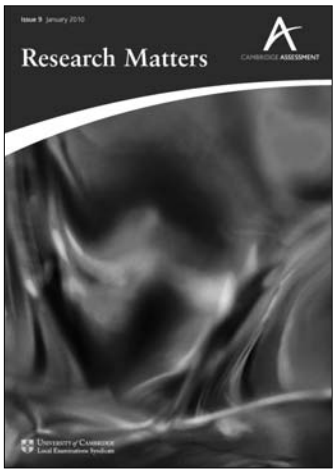
Issue 9 January 2010

Research Matters

CAMBRIDGE ASSESSMENT

University of Cambridge
Local Examinations Syndicate



If you would like to comment on any of the articles in this issue, please contact Sylvia Green.
Email:
researchprogrammes@cambridgeassessment.org.uk

The full issue and previous issues are available on our website:
www.cambridgeassessment.org.uk/ca/Our_Services/Research

Research Matters : 9

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

Once again, *Research Matters* is testimony to the breadth of focus and method which is present in research across Cambridge Assessment. The editorial rehearses the detail of the wide variety of approaches taken but what readers might miss is the crucial nature of the statistical reports listed at the rear of this issue. These are all available on the Cambridge Assessment main website and, in my view, make an essential contribution to the transparency of the examinations system. For a few years, QCA published time series data on GCSE and A level/AS results – the grade profile for almost all subjects, across all boards, in aggregate and also broken down by gender. QCA stopped publishing these in late 2006. These simple data – the numbers gaining the respective grades each year – are important enabling exploration of so many key issues: trends in the numbers getting A and A* grades; the low number of females doing some subjects (notably physics); and so on. Simple yet vital as we need to understand trends, and have comparable year-on-year data. Of course, they highlight some difficult issues and thus enable some hard questions to be asked. This, I believe, is a good thing. A better 'base level' of information allows journalists to explore issues with greater precision and understanding, and I would hope that young people, parents and other groups will be able to understand better the complex system we call 'education and training' and engage with the qualifications system on a more informed basis. The removal of these time series data from the QCA website marked a retrograde step in public accountability. By restoring these data to the public domain and presenting them in a clear and simple format, Cambridge Assessment has sought to increase transparency and make a genuine contribution to public accountability.

Tim Oates *Group Director, Assessment Research and Development*

Editorial

In this issue the themes range from the awareness and usage of non-standard English among sixteen year olds to examiner judgements in awarding. In the opening article Beth Black adds to the empirical research base on non-standard written English among young people at GCSE level. A questionnaire/assessment instrument was used to explore students' awareness of non-standard English and to investigate differences between school types, gender differences and regional differences.

Shaw and Cooke take us from 1858 to the present day with an analysis of history question papers for 16 year olds. They also used a variety of archive material to show more general developmental changes to the curriculum throughout the period. This article gives an interesting perspective on changes over time in both the structure and language of the papers as well as the marking schemes and processes of assessment.

The next three articles focus on expert judgements and the methods that impact on them. Novaković and Suto investigate the reliabilities of three potential methods for capturing expert judgement. These include traditional awarding (currently used), Thurstone pairs and rank ordering. A three-way comparison of the intra-method and inter-method reliabilities of all three methods was conducted in the context of setting grade boundaries. This research provides some interesting insights into the different judgemental methods, one of which is the current method used and two which could be useful in the future. In her article on how examiners make judgements about standards using different methods Greatorex used 'think aloud protocols' which involved Principle Examiners verbalising their strategies. The qualitative think aloud data were analysed using a framework designed for the purpose. This article offers many insights into what Principle Examiners attend to when they make judgements about grading standards.

In the final article Bramley discusses the theoretical rationale for using item level data in awarding. He presents some possible formats for displaying data and suggests ways in which the data could be used in practice. Data on individual questions or question parts are collected automatically as examination papers are scanned and marked on screen. These new processes provide a wealth of data that can be used to investigate how items function and how key discriminators can be used in awarding processes.

The Statistical Reports Series provides statistical summaries of a range of information using national-level examination data. The Factsheets are designed to make our research accessible to a wider audience 'headlining' main findings. Full reports can be found in the 'Conference Papers' section of our web-site.

Sylvia Green *Director of Research*

Exploring non-standard English amongst teenagers

Beth Black Research Division

Introduction

The main aim of this research is to measure the levels of awareness of non-standard English amongst GCSE level students.

There is a reasonable consensus on the conception of Standard English (SE) – a dialect or variety of English, (though with no local base). It is the most prestigious form of the language, its identifiable features residing in its grammar, vocabulary and orthography[1], but not in accent or pronunciation (Crystal, 1997, Trudgill, 1999). It is the variety of English used as the norm of communication in official communications, publications and broadcasting (Huddleston and Pullum, 2002). Paradoxically perhaps, as a spoken form, SE is probably a minority variety. Although it is widely understood, it is not widely used in spontaneous speech.

Non-standard English (NSE) is not just language which is merely different from SE, an accidental or one-off 'slip'; NSE implies a systemic feature of language which is shared with other speakers of the language but which diverges from the standard form. Nevertheless, an NSE utterance may have no marked semantic differential from the SE form. As Deborah Cameron (1995) points out, non-standard forms do not interfere with intelligibility – listeners are not bamboozled when decoding the intended meaning of Mick Jagger's 'I can't get no satisfaction'.

There is not a single NSE (in the way that there might be considered a single Standard English), but rather a number of forms, which include double negatives ('I can't find no money') and non-standard past simple or past participles (e.g. 'She brung me a drink').

The National Curriculum requires that pupils should be taught about differences between standard and non-standard English, and in particular, 'to be aware that different people make different choices about when standard English is appropriate' (DfEE, 2001). Certainly, the ethos of SE in the National Curriculum is intended to be descriptive, rather than prescriptive (Hudson, 2000), that English pupils should have a respect for their own and others' dialects and a sense of linguistic appropriacy – being able to shift their language usage in terms of register and form according to the communication situation. However, the expectation is that pupils will be able to write 'sustained standard English with the formality suited to reader and purpose' (Assessment Focus for Key Stage 3). There has, of course, been considerable and heated debate over the last two decades concerning the place of SE in the National Curriculum and how non-standard varieties or dialects should be treated (e.g. Cameron, 1995; Honey, 1997). This debate continues (e.g. QCA, 2005).

English is constantly evolving. Certainly, we are all familiar with changes in lexical usage as new words fill lexical gaps, find their way into usage and published dictionaries; slang terms become acceptable for formal usage. Change also takes place at the level of syntax[2]. Crystal illustrates this point by describing how the SE of today is not the SE of Jane Austen (Crystal, 1995). Differences include tense usage (*So, you are come at last*), irregular verbs (*so much was ate*), articles (*to be taken into the account*), prepositions (*she was small of her age*). It is an interesting point for prescriptivists to note, as such structures might now be considered as much non-standard as archaic.

There has been, however, less research into actual levels of usage of non-standard English. Hudson and Holmes (1995), investigating spoken English, found that about 30% of a selection of school children could speak for several minutes without using any NSE forms. Since this was produced in a rather formal school context it probably sets the upper limit (Hudson, 2000).

QCA (1999) found surprisingly little non-standard English in whole GCSE scripts, with 67% not displaying any non-standard forms.

Recent Cambridge Assessment research (Massey, Elliott and Johnson, 2005), using a cross-longitudinal design, identified a notable increase in non-standard usage in a sample of GCSE English scripts (stratified by grade) between 1980 and 2004, and in particular, between 1994 and 2004. The report also suggested that boys were more likely to use non-standard English forms than girls. Furthermore, as found in the QCA study, there was an indication that non-standard English usage was more prevalent amongst lower grades. The scope of this research, however, does not record the usage of the various NSE forms.

Lockwood (2006), in a cross-longitudinal study of 10–11 year-olds, found 'an overall decline in the children's awareness of standard English features' between 1999 and 2005, though this pattern was not uniform for all non-standard forms. Similar to Massey *et al*, he too found a gender difference, with males less likely than females to display awareness of NSE forms.

This project, through use of a questionnaire/assessment instrument, seeks to add to the empirical research on non-standard written English in young people at GCSE level. It attempts to add to the research in the following areas:

- which NSE forms are most and/or least recognised
- whether respondents could produce SE versions of the NSE forms
- whether respondents could spontaneously use the term non-standard English when asked to identify the type of English used in the assessment instrument
- perceptions of NSE
- whether characteristics of respondents (gender, school type, region) produce any differences in recognition and production of NSE

Method

To answer the above research questions an assessment instrument was devised in order to survey GCSE level students.

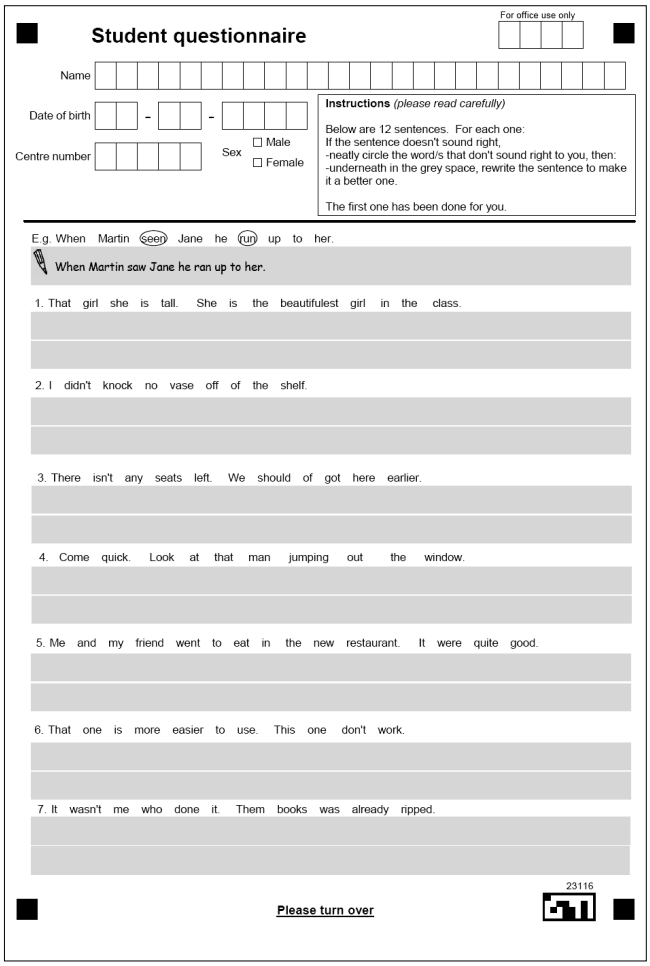
1 The accepted system of writing a language, including spelling and punctuation.

2 The rules and principles that govern the sentence structure of a particular language.

The NSE Assessment Instrument

The main part of the assessment instrument used in this study was broadly based upon that of Michael Lockwood's (2006) task. It contained twelve sentences/lines, each of which contained one or more NSE forms (see Appendix B). The sentences deliberately contained standard or even quite basic vocabulary in order to reduce the likelihood of adding an irrelevant source of difficulty.

The NSE forms contained within the instrument are shown in Figure 1. Not all the 41 NSE forms identified in a literature review were included, for reasons of brevity or the difficulty posed in creating a stand-alone sentence with an unequivocally NSE form. Finally, some NSE forms were deemed to be associated exclusively with particular dialects and these too were excluded.


Student questionnaire

For office use only

Name

Date of birth – –

Centre number

Sex ☐ Male ☐ Female

Instructions *(please read carefully)*

Below are 12 sentences. For each one:
If the sentence doesn't sound right,
-neatly circle the word/s that don't sound right to you, then:
-underneath in the grey space, rewrite the sentence to make it a better one.

The first one has been done for you.

E.g. When Martin (seen) Jane he (run) up to her.

When Martin saw Jane he ran up to her.

1. That girl she is tall. She is the beautifulest girl in the class.

2. I didn't knock no vase off of the shelf.

3. There isn't any seats left. We should of got here earlier.

4. Come quick. Look at that man jumping out the window.

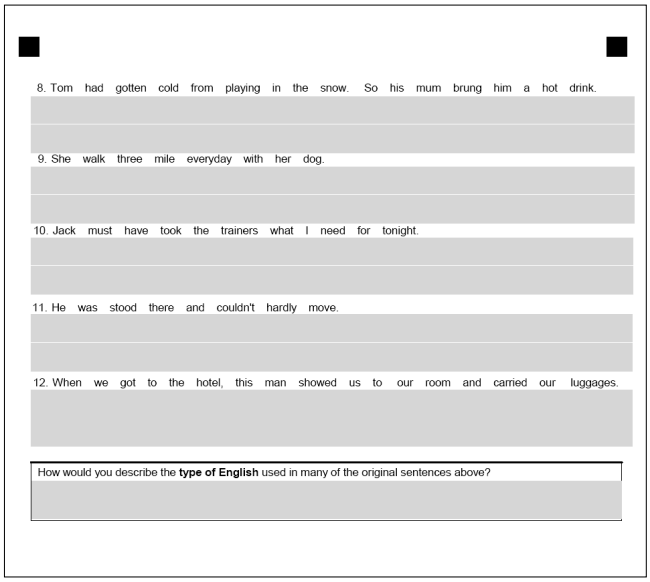
5. Me and my friend went to eat in the new restaurant. It were quite good.

6. That one is more easier to use. This one don't work.

7. It wasn't me who done it. Them books was already ripped.

23116

Please turn over


8. Tom had gotten cold from playing in the snow. So his mum brung him a hot drink.

9. She walk three mile everyday with her dog.

10. Jack must have took the trainers what I need for tonight.

11. He was stood there and couldn't hardly move.

12. When we got to the hotel, this man showed us to our room and carried our luggages.

How would you describe the **type of English** used in many of the original sentences above?

Figure 1: The Non-Standard English Assessment Instrument

Three NSE forms with origins in other 'Englishes' were included: 'gotten' – which is an American variant of the past participle of 'get'; noun phrase plus pronoun as subject ('that boy he went…'); and 'luggages', (treating an uncountable noun as countable), a common feature of second language English speakers in India, Singapore and Nigeria (Crystal, 1995). In all, 25 different NSE forms were tested on the assessment instrument.

Respondents were instructed 'neatly circle the word/s that don't sound right to you, then, underneath in the grey space, rewrite the sentence to make it a better one'. Thus, this provided both a test of recognition and production.

Respondents were also asked to describe the *type of English* used in these sentences. It was hoped that this would reveal something about the perceptions of NSE and whether or not respondents would be able to spontaneously produce the term 'non-standard English'.

There were other aspects of the research that will not be reported in this article for reasons of brevity. These included: (i) the measurement of students' perception of the varying appropriateness of NSE forms for communication contexts of varying formality; (ii) the administration of the same NSE assessment instrument to teachers and (iii) the analysis of the responses to a teachers' questionnaire about their attitudes towards teaching about SE and NSE in the classroom.

Sample

The sample consisted of 2098 students enrolled on English GCSE courses, of which 58.2% were male and 40.1% female (1.1% unrecorded). The students were from 26 schools, representing 23 different counties in England[3]. Although the original invited sample had been carefully constructed in order to represent the overall population in terms of geographic spread and centre type variation, the final sample that took part was more heavily weighted towards the independent sector (52.2% of respondents) than the general school population.

Materials

Each centre which had agreed to participate was sent multiple copies of the questionnaire (assessment instrument) so that there was one for each student enrolled on a GCSE English course (in either Year 10 or Year 11 in most cases). In addition, the contact teacher (usually the Head of Department) was also sent instructions to help them administer the questionnaire as well as standardised instructions to read out to the class. In brief, respondents were informed that the purpose of the research was to develop a national picture of English usage in England. They were instructed on how to complete the questionnaire and that there was not necessarily a single correct answer – they were asked to indicate what they thought was best or most appropriate.

The assessment instruments were completed in Spring Term 2007, or soon after the Easter holidays. In most cases, this represented the term immediately before the respondents completed their GCSE English course.

3 These were: North Yorkshire, West Yorkshire, County Durham, Greater London, Bedfordshire, Essex, Suffolk, West Midlands, Cambridgeshire, Derbyshire, Nottinghamshire, Shropshire, Merseyside, Tyne and Wear, Somerset, Devon, West Sussex, East Sussex, Wiltshire, Oxfordshire, Kent and Gloucestershire.

Analysis

The emphasis of this research was on particular forms of NSE and whether some forms were more readily recognised than others. There were multiple strands to the analysis:

- Rates of correct recognition for each NSE form.
- Rates of 'correct' versus 'incorrect' production for each NSE form.
- Overall 'scores' for recognition – each response was coded 0 (not recognised) or 1 (recognised) and totalled to give a score out of 25 for each candidate.
- Overall scores for production – each response was coded 0 (SE form not provided) or 1 (SE form provided) to give a score out of 25 for each candidate.
- Overall performance of the cohort on the assessment instrument including gender, school type and regional differences.
- Content analysis of responses to the question 'How would you describe the **type of English** used in the original sentences above'.

In order to reliably code whether the new version of the sentence was SE, three judges independently coded each response type during the course of an extensive content analysis. A discussion took place on all responses where there was not 100% agreement. In the majority of cases, this achieved a resolution. In about four cases where there was some disagreement, a fourth judge acted as the arbiter.

Results

Recognition and production rates

This analysis looked at which forms were most and least recognised, as well as the production rates – whether respondents could produce acceptable SE versions of the target NSE form (regardless of any other changes that might have been made which may have introduced a spelling error or even a non-target NSE form elsewhere in the sentence).

Table 1 shows a high correlation between NSE recognition and NSE 'correction', which provides some evidence of cross-validation of the two measures. However, perhaps counter to intuition, the rates of recognition (in all but one instance), are lower than that of correction. Possible reasons for this might include:

i. Respondents could not be bothered or did not realise they had to circle the relevant words even though they had recognised the presence of a non-standard form. There is some evidence for this as over 50% of respondents who did not correctly recognise a single NSE form had scores of over 20 in terms of correctly producing SE versions of the NSE forms.

ii. Respondents had either overlooked or had not consciously realised some of the NSE forms (because they did not interfere with comprehension), though naturally altered them at the point of production. This certainly seems likely in both items 8 and 19 where a quick read may not always register the missing –s or the missing preposition, though it is not a form the respondent would naturally produce.

Table 1: NSE forms according to most and least recognised and production rates of appropriate SE version of the target NSE form

Item	*NSE form*	*Example in NSE instrument*	*Recognition of target NSE form*		*Production of SE version of target NSE form*	
			Rank	*% recognised*	*Rank*	*% corrected*
1	Noun phrase + pronoun	That girl she is tall	4	88.0	5	93.4
2	-est form with adjective > 2 syllables	The beautifulest	7	84.2	10	87.6
3	Double negative	I didn't break no vase	1	91.2	4	95.3
4	Use of additional preposition	Off of	23	56.6	21	67.8
5	There is + plural	There isn't any seats left	14	74.0	20	69.9
6	Could of/should of etc	We should of	19	69.2	18	77.7
7	Use of adjective as adverb	Come quick	25	41.9	25	44.8
8	Loss of preposition	out the window	20	65.3	14	81.9
9	Me and xx as compound subjects.	Me and my friend…	24	53.8	24	56.4
10	Third person singular + were	It were quite good	3	88.8	1	97.2
11	More with +er	More easier	17	71.3	16	80.7
12	Third person singular + don't	That one don't work	8	83.9	6	92.6
13	Past Participle instead of past simple	It wasn't me who done it	13	75.9	13	83.5
14	Them + plural noun	Them books	10	79.6	11	86.8
15	Plural subject + was	[Them]…books was already ripped	11	78.7	7	92.4
16	Non standard past tense	Tom had gotten cold	12	77.5	15	81.7
17	Non-standard past tense	His mum brung him	5	86.6	9	89.2
18	Lack of subject verb agreement	She walk…	2	89.1	2	96.7
19	Measure nouns without plural marker	…three mile	9	79.9	3	95.8
20	Past simple instead of past participle	Must have took	16	72.4	19	77.0
21	Use of 'what' as relative pronoun	…the trainers what I need	6	84.7	8	91.5
22	Was sat/was stood	She was stood	22	57.3	23	58.2
23	Negative plus negative adverb	..and couldn't hardly move	18	71.0	17	78.5
24	This + noun to indicate newly introduced thing	This man showed us	21	58.8	22	65.5
25	Plural uncountable noun	Luggages	15	73.7	13	83.9

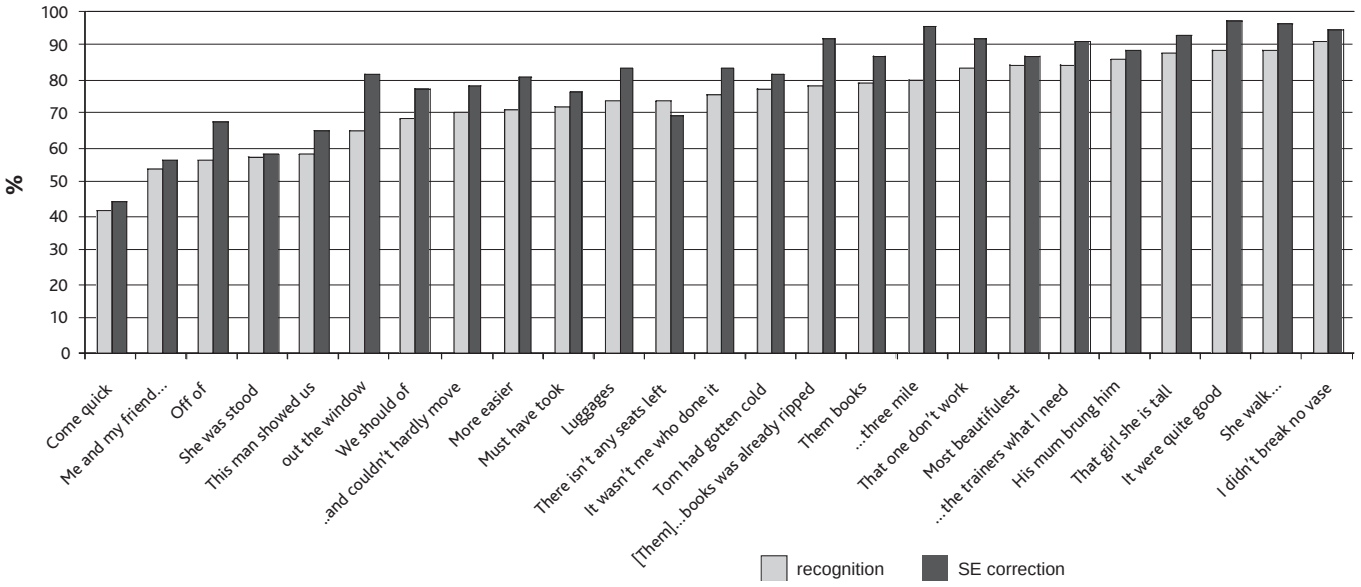


Figure 2: Recognition and production rates for each NSE form on the assessment instrument, arranged in ascending order according to SE production rates for NSE form

The only NSE form which bucked the trend and had higher recognition rate than production rate was 'there is' + plural noun ('There isn't any seats' – Item 5). In this case, some respondents who had circled the target NSE form struggled to produce SE versions.

Overall, the most commonly recognised NSE forms were the double negative (I didn't break no vase' – Item 3) the loss of inflection from 3rd person singular verb ('she walk' – Item 18) and subject-verb agreement ('It were quite good…' – Item 10).The most commonly successfully corrected forms also included 'she walk…' (item 18) and 'It were quite good…' (item 10) as well as 'three mile' (item 19). Certainly, recognition of double negatives and subject-verb agreement are flagged up in the National Curriculum as examples of non-standard English and this may explain the higher awareness in the respondents.

Non-standard forms which were least recognised and corrected were the use of adjective as adverb (omission of adverbial form -ly) as in 'Come quick' (Item 7) and the use of compound subjects 'Me and my friend' (Item 9). Interestingly, while some authors note that 'me and my friend' is 'unquestioningly non-standard' (Huddleston and Pullum, 2002), it is fairly standard in teenagers' conversation. 'She was stood' (item 22), 'This man..' (item 24) and 'off of' (item 4) all have less than 70% recognition and correction rates.

It is possible that these less well recognised NSE forms will find their way into SE, especially given the view that teenagers are linguistic innovators who bring about change in standard dialect (Kershwill and Cheshire, *in prep*).

Cohort profile

It is of some interest to see the distribution of respondents' scores on the questionnaire. It will give us some insight into how capable the cohort was overall at 'correcting' NSE forms. For the frequency graphs in Figure 3 the 'production' figures were used, rather than the recognition figures as these possibly represent possibly more sensitive outcomes.

The negatively skewed distribution (see Figure 3a) indicates that, overall, the cohort was quite capable at producing SE versions of target NSE forms as well as recognising them (see mean and modal scores in Table 2).

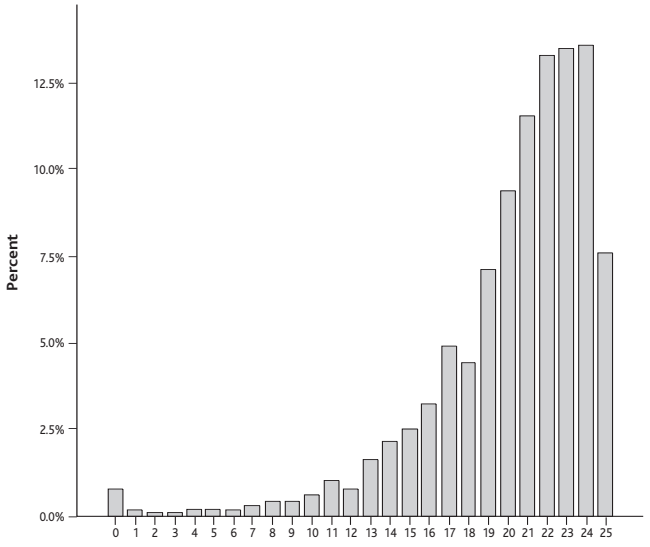
The difference between males and females in the 'production' score is significant, with females scoring slightly higher, though it must be noted that a difference of 0.44 in the means is only equivalent to 0.11 in terms of effect size for production scores and that there is no difference for recognition scores.

The comparison between state and independent schools reveals a highly significant difference, both with large effect sizes of 1.01 and 0.92 for production and recognition scores respectively[4]. While the difference between state maintained and independent sectors is significant and in favour of the students in independent schools, it is not possible to determine the cause of this difference within this study (e.g. academic ability, educational experience etc.).

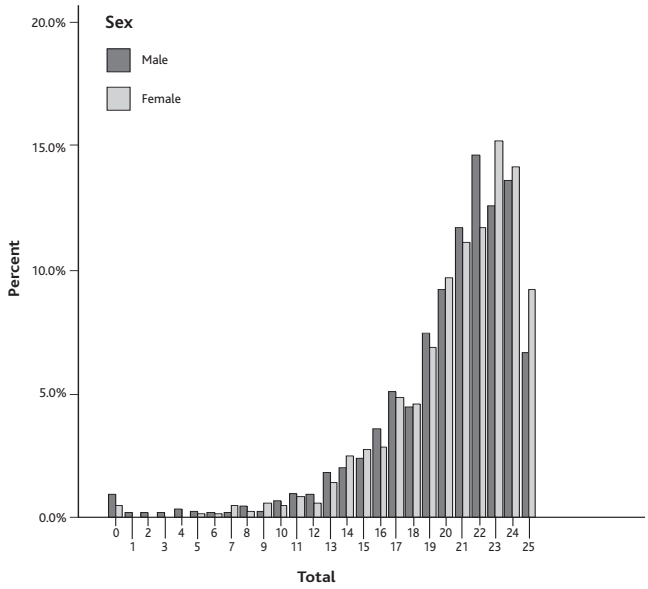
The difference between north and south is also significant. While references to the 'north-south divide' are common in geographical, political and economic discourse, there is no universally agreed, single and exact North-South dividing line. Rather, it moves according to various indicators (Green, 1988). In this research, the line was determined by appropriate groupings of the nine English 'NUTS 1' areas (Nomenclature of Units for Territorial Statistics) as used by the Office of National Statistics. The dividing line runs roughly from the Severn Estuary to the Wash (Figure 4). This North-South dividing line is not dissimilar to the geographical line which divides upland from lowland England. Respondents were grouped according to the location of their centre[5]. In this research, respondents from the northern counties obtained a higher score on average. This is possibly counter intuitive: although Standard English is not associated with any specific local base or dialect, there is a common perception that it is something more associated with the south. The effect sizes are moderately small – 0.20 for both production and

4 Effect size calculated using the version of Cohen's d where the denominator is the pooled standard deviation (Cohen, 1988). Effect size takes account of the magnitude of difference between the groups. Unlike significance tests, effect size is independent of sample size.

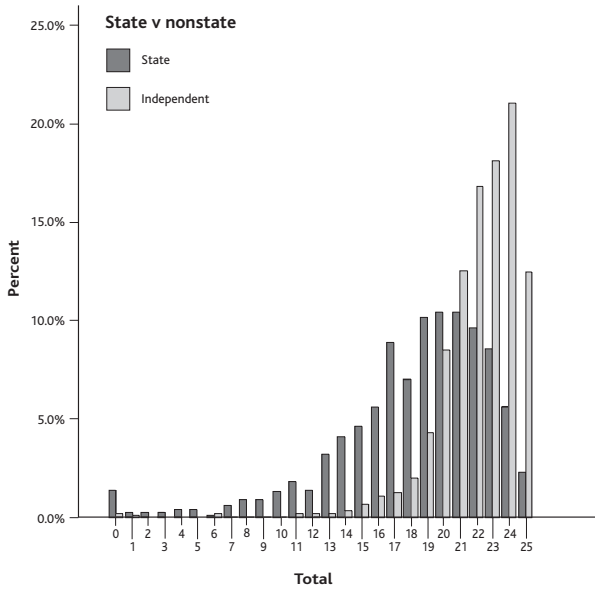
5 In the vast majority of cases, candidates' place of residence is likely to be in the same NUTS area (and therefore any larger regional grouping) as that of their school. However, in the case of independent schools, especially if they are prestigious, candidates may live much further afield.



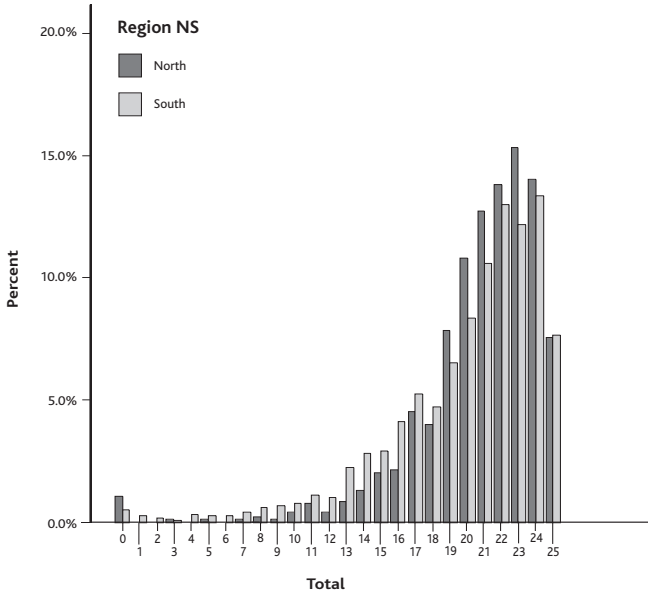
3a) Overall distribution of 'production' scores



3b) Distribution of 'production' scores by gender



3c) Distribution of production scores by school type[6]



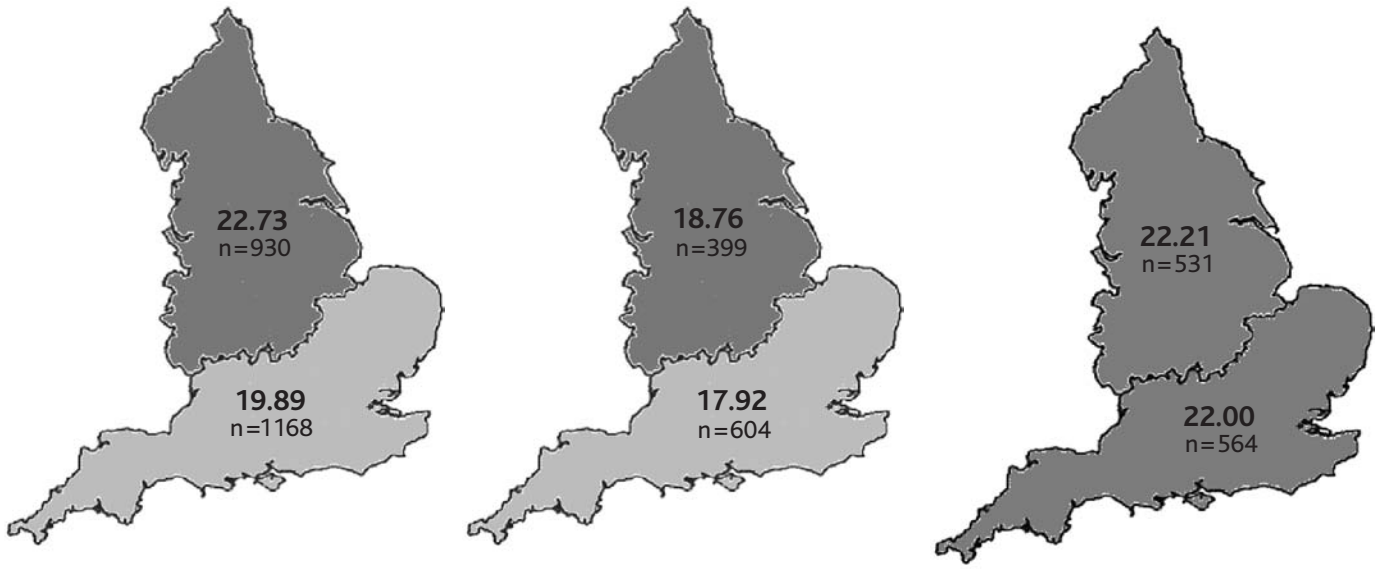
3d) Distribution of production scores by region (North/South)

Figure 3: Distribution of production scores on the assessment instrument for the whole cohort and with breakdowns for gender, school type and region

Table 2: Descriptive statistics for distribution of 'production' and 'recognition' scores on the assessment instrument

		Production scores					*Recognition scores*				
		Mean	*s.d.*	*Median*	*Mode*	*Sig p =*	*Mean*	*s.d.*	*Median*	*Mode*	*Sig p =*
	Overall	20.26	4.26	21	24		18.55	5.71	20	22	
sex	Male	20.13	4.38	21	22	0.020	18.61	5.65	20	22	ns
	Female	20.57	3.91	21	23		18.60	5.73	20	22	
school type	State	18.25	4.76	19	20 & 21	0.000	16.05	6.18	18	19	0.000
	Independent	22.11	2.65	23	24		20.85	4.07	22	22	
region	North	20.73	3.85	22	23	0.000	19.19	5.44	21	22	0.000
	South	19.89	4.53	21	24		18.05	5.88	20	22	

6 For this bar chart, the state school category includes comprehensive and sixth form respondents.

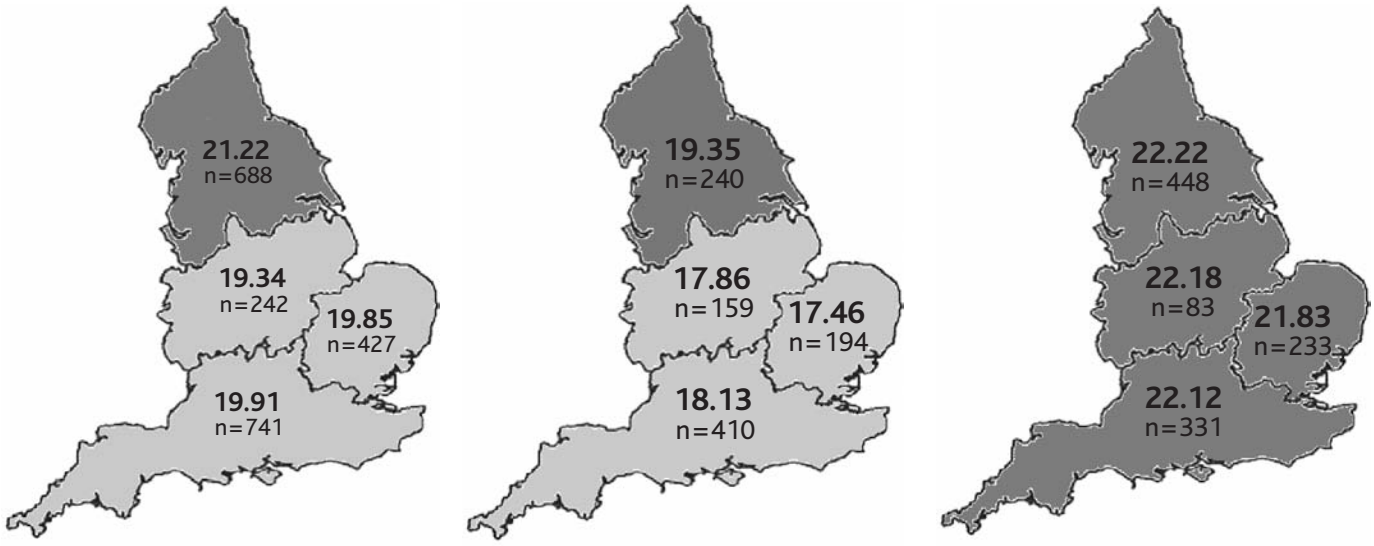


4a) All respondents

4b) State school respondents only

4c) Independent school respondents only

Figure 4: mean production scores by region – North, South



5a) All respondents

5b) State school respondents only

5c) Independent school respondents only

Figure 5: mean productions scores by four regions – North, Mid, East and South

recognition. In part, it might be thought that the better scores for the candidates from the North is because there is a disproportionate percentage of independent school candidates compared to the south. However, when the mean for north and south is calculated for independent schools and state schools separately (see Figure 4), we can see that there is almost no difference between north and south for independent school students, but a difference for state schools (p=0.000). This pattern is replicated also when region is looked at in terms of four areas – North, South, Mid and East – see Figure 5 – whereby we can see an overall significant difference between the mean production score (p=0.000), no difference for independent schools (p=0.307), but a significant difference for state schools (p=0.000). In some ways, it is unsurprising that we can see more variability for state school respondents by region, when reminded of the overall distributions of the two populations (see Figure 3c and Table 2).

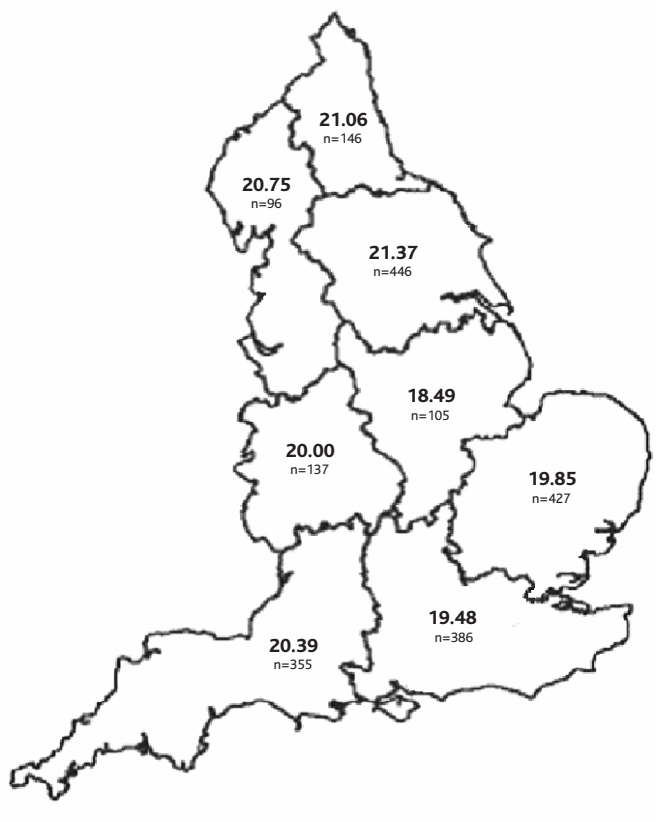
Finally, regional differences in overall production scores were analysed by grouping respondents' centres according to NUTS areas (see Figure 6). It is worth noting that, when looking at these smaller regions, perhaps only 3 or 4 schools might be in any one region and it is unknown how representative the schools' intake and production of NSE might be of any specific region. Any differences between regions may be just an artefact of the data rather than any real effect.

For Figures 4 and 5, differences in the shading of areas indicates a statistically significant difference between the regions.

Perceptions of NSE

Respondents were asked to name the type of English in the sentences in the instrument. One aim of the National Curriculum for literacy is for students to be able to identify standard versus non-standard English, and that they should also see NSE as a dialect with equivalent status to SE,

Figure 6: Mean production score by region – NUTS areas



North East	Northumberland, Tyne and Wear, County Durham, Tees Valley	21.06
North West	Cumbria, Lancashire, Merseyside, Greater Manchester, Cheshire	20.75
Yorkshire and the Humber	North Yorkshire, West Yorkshire, East Yorkshire, East Riding, North and North East Lincolnshire	21.37
East Midlands	Derbyshire, Nottinghamshire, Lincolnshire, Leicestershire, Rutland and Northamptonshire	18.49
West Midlands	West Midlands, Herefordshire, Worcestershire and Warwickshire, Shropshire, Staffordshire	20.00
East of England	Norfolk, Suffolk, Cambridgeshire, Bedfordshire, Hertfordshire, Essex	19.85
London and the South East[7]	London all boroughs, Berkshire, Oxfordshire, Buckinghamshire, Hampshire, East and West Sussex, Surrey, Kent	19.48
South West	Gloucestershire, Wiltshire, Somerset, Dorset, Devon, Cornwall.	20.39

though not appropriate for formal spoken or written English. It was hoped that this question would give some insight into perceptions of NSE.

Many respondents included more than a single codable response such as:

Respondent #962: *wrong/slang/improper English*

Respondent #1656: *Coloquel and like they speak in Eastenders! Informal, conversational*

Respondent #1390 *Common, peasantry, 'chav', Incorect, Heinous grammatical errors*

Respondent #1380 *Formal and third person*

Responses to this question were coded according to the first codable unit in any response as any coding method to take account of combinations of descriptions involved in excess of 100 categories. Thus, the examples above would have been coded, in turn, as 'wrong', 'colloquial', 'common', 'formal'. Figure 7 indicates the frequencies of the first codable unit in any response.

Figure 7 shows that the four most common first responses (according to codable response) were 'slang', 'informal / casual', 'colloquial' and 'bad / poor'. A number of respondents identified the language as 'childlike' (or 'like a 5 year old') – and it is possible that for these respondents the salient features of the language were *not* the non-standard forms per se, but the simple sentence structure and vocabulary.

Overall the term non-standard English (or 'not standard English') was present (at any point) in just 2.8% of responses (n=59) (compared with 3.4% (n=72) for 'chavvy'). Thus, it seems that most of the respondents could not spontaneously deploy the term non-standard English.

There were some respondents who identified the type of English as a specific dialect (see Figure 8).

Interestingly, these identified dialects range (in addition to American) from north-east to south-west England, and in the majority of cases represent a geographic locality close to the respondent. In these cases, it is not always possible to know whether the respondent themselves identified with a specific dialect (their own in-group), or regarded it as belonging to an out-group.

Further analysis was required to discover whether respondents viewed NSE (regardless of whether they had used this term or not) as of equal status to SE as is the intention of the National Curriculum, or as a lower status form. This involved recoding the first codable units (as seen in Figure 7) into either 'neutral' or 'negative'. Thus, responses originally coded as 'colloquial', 'informal', 'casual', 'abbreviated', 'teen speak', 'everyday' were coded as neutral; while 'bad', 'poor', 'disgraceful', 'pikey', 'Pidgin' (NB: in every case, spelt like the bird), 'unintelligent' and so on, were coded as negative. The percentages are presented in Figure 9.

Overall, (see Figure 9a) respondents were more likely to present the NSE forms as negative/inferior than give a response indicating a more neutral stance[8]. There was little evidence of any gender difference in these perceptions, or, as one might have imagined, a state versus independent school difference. However, there was a difference in terms of region (North/South) which was significant ($p=0.000$[9]). From the graph (see Figure 9d) we can see that overall the respondents from the northern counties were less likely than the respondents from the southern counties to hold a negative view of NSE. This finding, together with higher production and recognition scores for northern respondents, may indicate that these respondents have more readily absorbed the values of the National Curriculum towards SE/NSE.

Other responses, which provide some insight into attitudes and understanding of language, include:

Respondent #2017: *Confused tenses (a.k.a. Russell Brand speak.) and plural adjectives and verbs incorrect. In a word 'childish.'*

Respondent #78: *I can describe this type of English like a type of simple language what we can use when we speak with friends.*

Respondent #1263: *COMMON/AGRICULTURAL*

7 For this figure, the NUTS areas of South East and London were merged because of the low sample size in one area.

8 It is likely that if the coding were based upon the whole response, that the proportion of negative responses would increase.

9 Determined using a Chi-Square test.

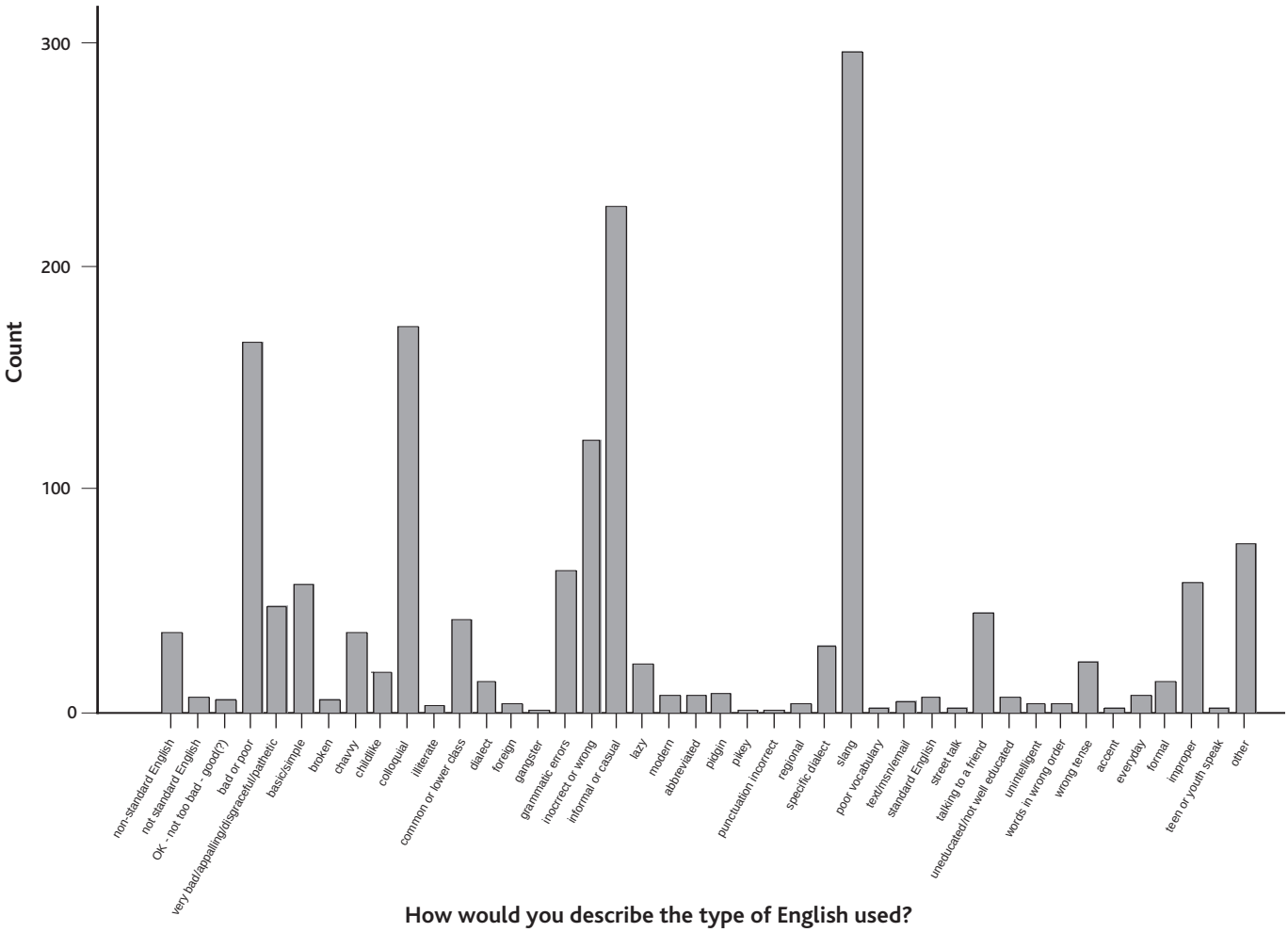


Figure 7: Frequency counts of first codable unit of responses to the question 'how would you describe the type of English used in the sentences above?'

Respondent #929: *It is understandable however there are many mistakes.*

Respondent #1456: *Not correct, yet understandible*

Respondent #1400: *Bristolian/chav*

Respondent #291: *incorrect, slang, use of double negatives*

Respondent #126: *disscorrectly ordered*

Respondent #1222: *Some of the original sentences had small mistakes and there were bit unproper.*

Respondent #301: *Standard english /poor grammer*

Respondent #447: *The original sentences have different dialects which make them incorrect*

Respondent #102: *Very informal, as you would talk to a friend or over an instant messaging programme (msn).*

Respondent #1898: *written in a Regional accent. Non standard english*

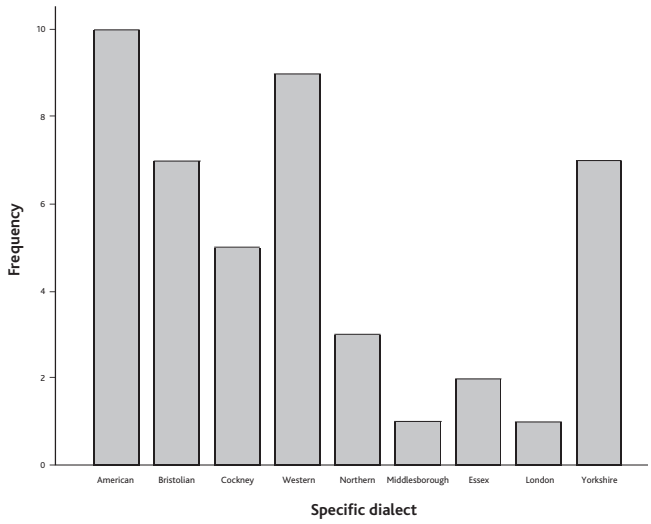
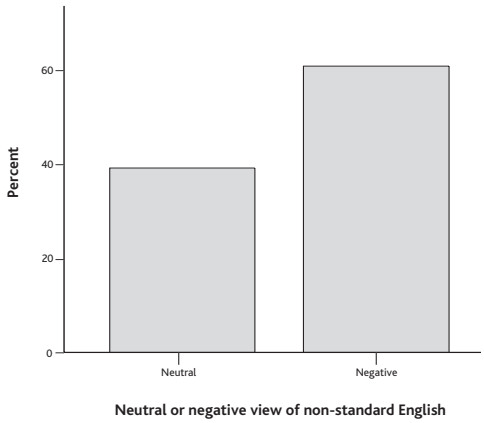
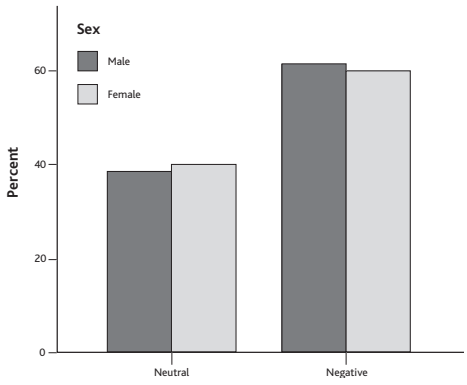


Figure 8: Frequency of identified specific dialects

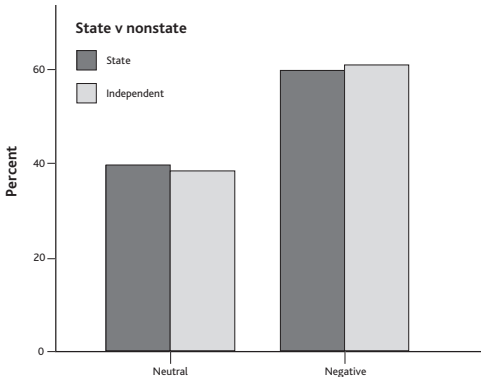


9a) Overall distribution of perception of NSE

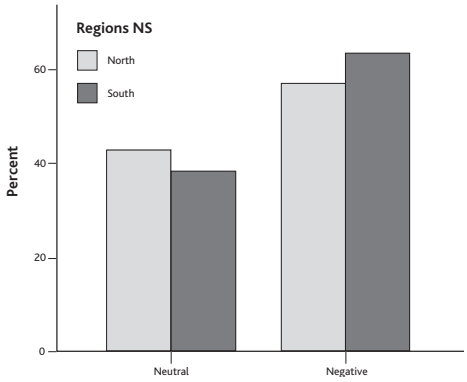


9b) Distribution of perception of NSE by gender

Figure 9: Perception of NSE of respondents, as coded on the basis of the first codable unit of response to the question



9c) Distribution of perception of NSE by school type[10]



9d) Distribution of perception of NSE by region

Discussion

There are some interesting similarities and differences between this research and that of Michael Lockwood, though it must be remembered that Lockwood's study looked at a younger age group. Similarities include:

- Some gender difference, though not large, in awareness of NSE. Lockwood's own longitudinal survey points to a closing gender gap as a result of declining female awareness rather than increasing male awareness.
- High awareness of the various NSE forms which involve subject-verb agreement.
- Similar rates of identifying 'gotten' as NSE (77.5% in this study versus 70% in Lockwood's).

Some of the differences are worth pondering. One might speculate whether the differences are due to research design issues such as the choice of sentences, sample size, or age of the respondents. It is possible that children 'grow out of' some forms of NSE between the age bands of 10–11 and 14–16. These differences include:

- Adverbial use of adjective ('Come quick') was the least commonly recognised form in this survey, though one of the most commonly recognised in Lockwood's ('We done our work proper').
- In Lockwood's study, 'could of' was accepted by 92% or respondents as standard, averaged over the three sampling years. However, this study reports that only 20% of respondents failed to correct this form. This may suggest that this is one feature of English at which children improve with age.
- In Lockwood's study, 'Me and my dad' was accepted as SE by 86%, compared to 43.5% in this study, again, possibly indicative of awareness increasing with age.

Limitations

While this research had a very large sample, there were some limitations which included:

- The assessment instrument contained contrived sentences in order to try to produce clearly non-standard examples. Their contrived nature may not have been sufficiently convincing or life-like and may have confounded responses.
- Whilst this research shows that, for example, 'Come quick' (use of adjective as adverb) was the least commonly recognised and 'I didn't knock no vase' (double negative) as the most recognised, these results might not necessarily generalise to other examples of the same form such as 'I did it easy', 'speak proper' or 'I'm not never going back there again'. Different syntax and construction may alter the perception of a sentence or form within a sentence as non-standard.
- This research involves only written English, and did not tell us about the usage of these forms in spoken English.
- From this research alone, and without replication of this work in several years' time, it is not possible to know whether the usage and awareness of NSE is stable, increasing or decreasing.

10 For this bar chart, 'state school' includes both comprehensive and sixth form respondents.

Conclusions

In summary, this research indicates:

- On the whole, recognition rates of NSE and production rates of SE were quite high.
- Despite National Curriculum aspirations not to treat SE as the prestige version, the majority of respondents identified the language in the stimulus sentences as of an inferior type.
- There are significant differences in school types (independent versus state) in terms of correct production of SE versions of NSE forms.
- There is a small though significant difference between males and females in correct production of SE versions of NSE forms
- There is some evidence of regional differences in NSE production – in particular for a North-South divide.

References

Cameron, D. (1995). *Verbal Hygiene*. London: Routledge.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Crystal, D. (1997). *The Cambridge Encyclopaedia of Language*. Cambridge: Cambridge University Press.

DfEE (2001). *Key Stage 3 National Strategy Year 7 sentence level bank*. London: DfEE.

Green, A.E. (1988). The North-South Divide in Great Britain: An Examination of the Evidence. *Transactions of the Institute of British Geographers* **13**, 2, 179–198.

Honey, J. (1997). *Language is Power*. London: Faber and Faber.

Huddleston, R. & Pullum, G.K. (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hudson, R. & Holmes, J. (1995). *Children's Use of Spoken Standard English*. London: School Assessment Authority.

Hudson, R. (2000). *The Language Teacher and descriptive versus prescriptive norms: The educational context*. Accessed at http://www.phon.ucl.ac.uk/home/dick/SEhudson.htm on 07/06/2008.

Kershwill, P. & Cheshire, J. (in prep). Linguistic Innovators: The English of London Adolescents. Accessed at http://www.lancs.ac.uk/fss/projects/linguistics/innovators/overview.htm on 10/05//2007.

Lockwood, M. (2006). Changing Standards? Children's awareness and knowledge of features of written standard English at ages 10–11. *Changing English*, **13**, 1, 17–28.

Massey, A.J., Elliott, G.L. & Johnson, N.K. (2005). *Variations in aspects of writing in 16+ English examinations between 1980 and 2004: Vocabulary, Spelling, Punctuation, Sentence Structure, Non-Standard English*. Cambridge: Cambridge Assessment.

QCA (1999). *Technical accuracy in writing in GCSE English: research findings*. London: QCA.

QCA (2005). *English 21. Playback: a National Conversation on the Future of English*. London: QCA.

Trudgill, P. (1999). Standard English – what it isn't. In: T. Bex & R.J. Watts (Eds.), *Standard English: the widening debate*. London: Routledge.

ISSUES IN QUESTION WRITING

The evolution of international History examinations: an analysis of History question papers for 16 year olds from 1858 to the present

Stuart Shaw CIE Research **and Gillian Cooke** Cambridge Assessment Archives

Background

The focus of this article is on international History examinations for 16 year olds from 1858 to the present day and examines the historical/cultural context for, and the setting of, these examinations in the medium of English. Specific reference points throughout this period have been taken and a linguistic analysis applied to the question papers. A variety of archive material has been used to show more general developmental changes to the curriculum throughout the period. The article examines the language used, the candidate base, the regional differences of the papers and the examiner expectations. To put these findings into context, other sources, including examination regulations, examiners' reports and subject committee papers have also been studied.

In 1858 when the Cambridge Local Examinations were introduced, History was a compulsory element of the Junior examination. Candidates had to pass in a whole range of subjects to gain a school leaving certificate and English history could not be avoided. 150 years later there is no doubt that school examinations for 16 year olds have undergone radical transformation and for History examinations to have remained unchanged would be unthinkable. The interest lies not in the fact that the examinations have changed but in the way they have changed. While the trend is inevitably towards a more familiar, contemporary style, this study also shows that the pace and particular directions of change have been of a less predictable nature.

Challenges and constraints

The aim of the study is to determine how History examinations have evolved. The selection of History question papers from different periods in time should be based on some assumption that comparisons across time are on a 'like for like' basis. However, this was not found to be the case. The question papers are drawn from different examinations: the Cambridge Junior Local Examination until the end of World War 1,

the School Certificate from 1918, and the International General Certificate of Secondary Education (IGCSE) from 1988. There is every reason to expect discontinuities caused simply by changes to the examining system, though there are some notable exceptions. For example, School Certificate still exists as an examination, and History papers are set for it. In effect, there was no universal change after 1988. Rather, IGCSE was developed as an examination for a different target market. Similarly, all overseas centres did not simply continue with an unchanging School Certificate after 1951; rather, School Certificate evolved in a variety of ways to include aspects of the GCE O Level examination.

This raises a second issue – who were these examinations for? Can we at least argue continuity in this respect? In one sense the answer is yes. In the broadest possible way we can regard all three examinations as equivalent to an English 16 plus examination, the level at which some students might leave full-time schooling. However, more specifically there are differences. At first Cambridge Junior was taken only by 297 English boys. By the end of the nineteenth century there were a few candidates from overseas centres (about 370), but this made no difference to the nature of the examinations, and these candidates were largely sons and daughters of British colonial administrators.

After the First World War, the numbers of home and overseas candidates increased rapidly, along with an emerging awareness that an English examination, the School Certificate, might not be entirely suitable for non-British students. This led to the development of History question papers for specific areas, for example Indian History, which were not aimed at British candidates. However, it was impossible to set such papers for all areas so most overseas candidates still took exactly the same papers as British candidates. By then there were examinations twice a year, July for home candidates and December examinations for overseas candidates. This rather hybrid system came to an end in 1951 with the introduction of the GCE O Level. At first this applied only to UK candidates, whilst School Certificate continued internationally.

Between 1858 and 1951, then, the candidature of the December examinations evolved from being entirely British to entirely non-British. The great majority of these School Certificate candidates came from what, with the achievement of independence in former colonies, were known as 'Ministry' areas. In effect they were students from government schools in countries that chose to use Cambridge examinations. Throughout the 1950s regional history papers were developed through the new Regional Awarding Committees – West Africa, East Africa, the Caribbean, Mauritius, Malaya – more often than not the precursor to localisation projects aimed at countries wishing to establish their own examining boards.

By the 1980s, however, the need for further change was becoming apparent. Syllabuses in many subjects, including History, were becoming dated, and a new market of English-medium, non-Ministry schools was emerging. A new examination, the IGCSE emerged which incorporated the kinds of changes included in the GCSE examination introduced in the UK in 1988. These English-medium, international schools were of a markedly different nature to many Ministry schools. They were well resourced, willing and able to innovate, and with students drawn mainly from professional backgrounds whose English-language skills were good enough to cope with the demands of less traditional examinations.

The last issue is perhaps the most fundamental. It is hard to be certain about whether there is any continuity in what these examinations were setting out to assess as until surprisingly recently syllabuses in History did not include assessment objectives. Had you asked the examiners in 1858 what they were testing, they would certainly have replied "History". If you asked them today they would say something like "Historical knowledge and understanding, the ability to construct explanations, and the skills of handling historical source material". During this period there has been a huge change in what is understood as the study of History, and the examination papers reflect this. Even the most superficial scrutiny of the papers from 1858 and 2000 reveals the almost entirely different demands they make on candidates.

An associated problem is whether examinations, with or without assessment objectives, actually test what they claim to be assessing. The traditional criticism of History examinations was that, although they asked questions which seemed to demand explanations or analysis of historical events, they were in fact marked solely on the basis of knowledge of the events. Without marking schemes, it is hard to be certain of the justice of this claim.

Identifying a methodology for analysing question papers

Twelve History question papers were selected for the study and the period was divided in four:

1. Early Locals from 1858 to 1917
2. Late Locals/School Certificate from 1928 to 1951
3. Post 1951 to 1972
4. IGCSE, 1989 to 2000

A general overview of each period, drawing on Examination regulations and specifications, Examiners' reports and history committee files is followed by a question paper analysis. Analysis includes consideration of the lexical, structural and functional resources used; English provided in the question, in the rubric, the English expected of the candidates and the general instructions to candidates.

1. Early Locals

Overview

Initially, the candidates were all boys but the examinations were opened to girls on an equal standing from 1865 and the statistics show that the girls enjoyed considerable success from the start. The examiner for the Preliminary Cambridge Examinations for the English History paper commented in 1866 that 'the style of the girls' replies' was 'better than that of the boys. It was more straightforward and to the point, and there were fewer attempts at fine writing.'

The examiners' reports are not noticeably dated. The 29th Annual Report includes complaints about 'vagueness', 'inaccuracy', 'slavish reproduction of the words of text-books' and concludes, 'the best work was done by girls.' But this was written by examiners in 1887, who were predominantly Cambridge Dons and Clerics. The history examiners themselves generally came from the Classics and English disciplines, which makes particularly poignant the criticisms about the candidates' lack of historical perspective. Use of obsolete text books and regurgitation of facts rather than answering the question are also 'modern' criticisms which appear in 1899, challenging the notion that to pass history examinations during this period required only a knowledge of historical facts.

The examiners do not shy away from negative comments but nor do they lack humour. Mistakes in history have long been a potential source

of amusement and J N Keynes' commonplace book includes many comments from history papers. 'Henry VIII was a very waistful king' wrote one candidate in the 1880s. Another, in 1882, 'described Edward the Black Prince as having been present at Hastings, Agincourt and other battles ranging over a period of 300 years and wrote of him being just 16 years of age at the latest of these fights.' There was no discernible improvement during the period, for in 1915 the examiners of the Junior English History paper wrote: 'Many candidates exhibited a hopeless ignorance of chronology.'

At first, the English History part of the Junior Examinations was a compulsory element, along with arithmetic and dictation, but by 1874 History had become an optional part of the English section, naturally entitled English History. Candidates could choose between a paper on the History of England, Roman History, Geography or Shakespeare. During the 1890s English Grammar was introduced into the group and in 1899 a separate section for History and Geography emerged. Junior candidates thereafter could choose one history paper from History of England, Roman History or the new paper on the British Empire and could take this together with the Geography paper if they wished to take Group 3 examinations.

Take up of the optional English History paper among Junior candidates remained high even after the introduction of the British Empire paper. In 1899, over seven thousand candidates out of a total Junior entry of 8,277 took the English History paper and in 1915, fifteen years later, 9,302 candidates opted for English History and just 417 for the British Empire.

The examiners' reports on the British Empire paper are not particularly positive. In 1902 they commented that 'many of the Boys sent up almost worthless papers'. Overseas candidates, who were presumably more likely to take the paper, attracted little specific attention until later in the period but in 1913 received the following encouragement: 'Several colonial centres had evidently paid special attention to the history for that part of the Empire in which they are situated. This is an excellent plan; but care should be taken that it does not involve neglect of highly important occurrences in other parts of the Empire.'

Were these candidates local or the children of British colonial administrators? The candidate base is not clear as records of entries do not exist so available evidence shows only passes – many of whom appear to have been British expatriates. By 1917, the colonial candidates are of mixed origin and by no means uniform throughout the colonial centres. There are many English names on the pass lists for India, but comparatively few for candidates from Penang and Singapore. Although 'Colonial Centres' were sent their own regulation notices, the syllabus for all candidates, in History at least, remained the same.

The examination regulations for 1917 are remarkably similar to those of 1899 and the set texts books show that the periods selected for examination followed a rather predictable cycle alternating largely between the years 449–1509, 1509–1688 and 1688–1832; indicative of a traditional or unimaginative approach by examiners as well as thorough record keeping.

Question paper analysis

For any examiner with experience of marking a wide variety of late 20th century History examinations, these papers would seem the most distant and different in nature, reflecting a way of studying the subject that has now completely disappeared.

The earlier question papers seem most focused on factual recall – listing, naming, giving dates. In the later papers there is a noticeable move away from pure recall and towards a demand for explanation – or extended description, with more emphasis on opinion and scope for creativity. Candidates need to be able to produce complex sentences and longer, more cohesive text. Past simple, continuous and perfect tenses (active and passive) would be commonly used as would comparative forms. This is an interesting shift in how the nature of the subject must have been perceived.

The increase in the whole paper time allocations is also an indication that examiners sought more discursive answers. In fact, in these papers it is possible to discern the standard pattern for School Certificate History examinations of the next century beginning to emerge.

The choice of content reflects a mid-Victorian view of History as the study of English kings and queens with later additions of French Monarchy and Constitutional History. Content choice would subsequently emerge as a major issue in History syllabus development, sometimes dealt with by offering alternative papers, and sometimes by offering wide question choice within papers. The optional papers are interesting in showing a concern for Empire, either British or Roman.

Candidates would need a wide range of lexis to answer these questions. Political, legal and historic language might be required to describe methods of legislation, explain political questions, state the chief Privileges of Parliament, or to describe treaties, events and foreign policies. Lexis is not always selected for accessibility: for example, 'What was the issue of their attempts?' and 'the situation of the battle-field'.

The papers are presented in a very formal, impersonal style, the register being maintained by the use of passives and by addressing the third person not the candidates themselves. But there is a gradual change in register – instructions are worded as 'candidates may' as opposed to 'candidates are expected to'. The rubrics appear to become more accessible as they inform candidates that they 'must pass in both parts of the paper' as opposed to 'must satisfy the Examiners in this Paper'.

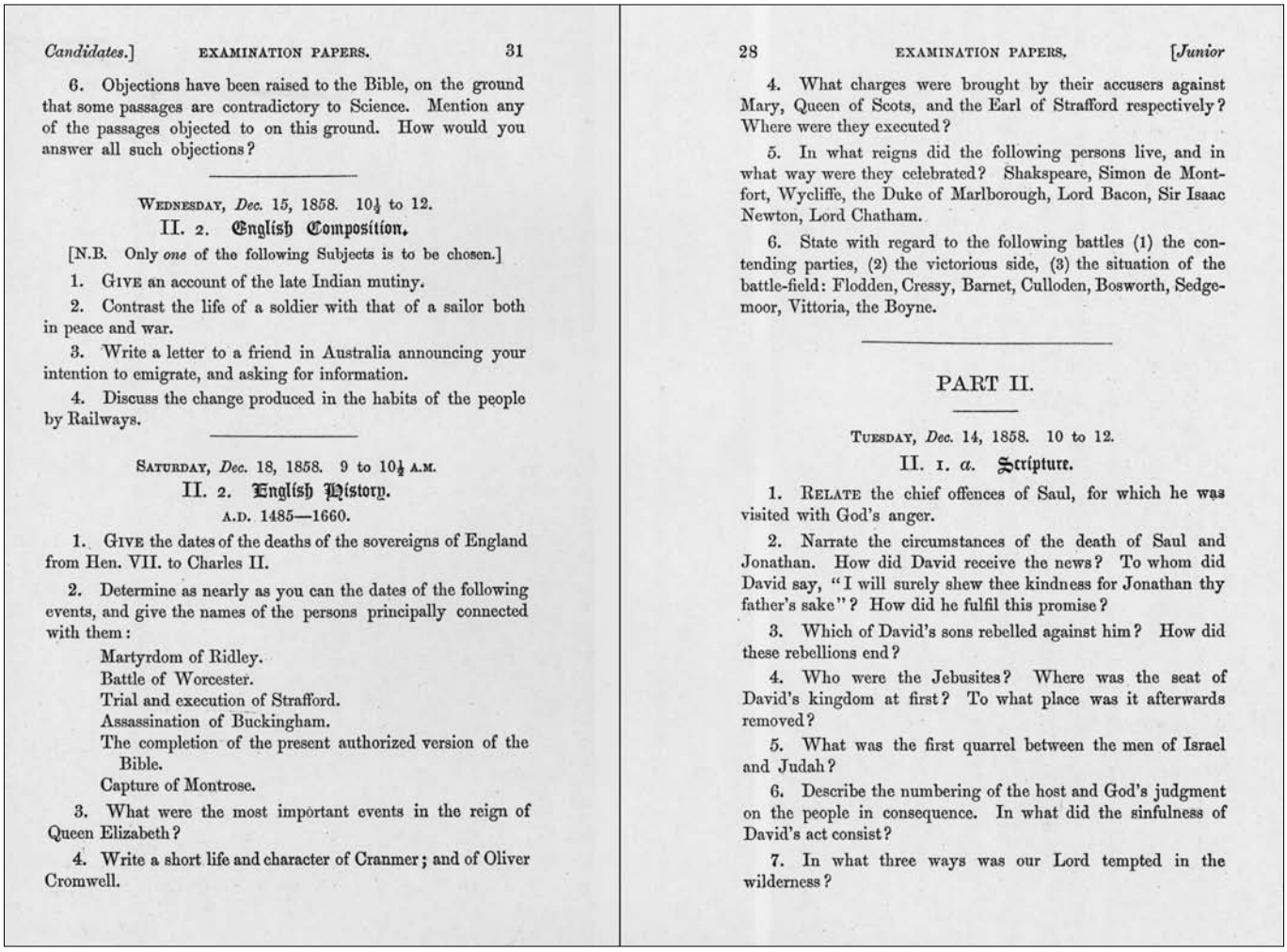
A greater range of functional language is used across the papers. There could be some duplication of meaning which might cause confusion with different verbs being used to express the same function. For example, verbs include 'describe', 'write a brief account/history of...', 'tell what you know of...', 'shew', 'discuss', 'compare', 'distinguish between', 'mention', 'set forth', 'set down', 'trace', etc. The question structure also changes over time as imperatives are used far less frequently and there are more past simple, present simple passive, and past simple passive questions.

There is no indication of expected output – in terms of length or style, mark allocation, or suggested timing per question. Lack of such information would not help candidates to perform to the best of their abilities in an examination situation. Despite this, the demands placed upon candidates across the papers appear to be similar.

2. Late Locals/School Certificate

Overview

By 1928 the School Certificate and Higher School Certificate had been offered to candidates for ten years and was well established as the first national school examination. As well as the School Certificate, UCLES still offered the Cambridge Junior and Senior examinations to overseas candidates, together with an impressive range of specialist regulations for particular overseas candidates; syllabuses, for example in Urdu and Hindi for Indian candidates. The Junior examination regulations for History remained as they were when they were introduced in 1899 but the school certificate candidates could choose between three different periods of English History or British Empire, Modern European, Roman or



Candidates.] EXAMINATION PAPERS. 31

6. Objections have been raised to the Bible, on the ground that some passages are contradictory to Science. Mention any of the passages objected to on this ground. How would you answer all such objections?

WEDNESDAY, *Dec.* 15, 1858. 10½ to 12.

II. 2. **English Composition.**

[N.B. Only *one* of the following Subjects is to be chosen.]

1. GIVE an account of the late Indian mutiny.
2. Contrast the life of a soldier with that of a sailor both in peace and war.
3. Write a letter to a friend in Australia announcing your intention to emigrate, and asking for information.
4. Discuss the change produced in the habits of the people by Railways.

SATURDAY, *Dec.* 18, 1858. 9 to 10½ A.M.

II. 2. **English History.**

A.D. 1485—1660.

1. GIVE the dates of the deaths of the sovereigns of England from Hen. VII. to Charles II.
2. Determine as nearly as you can the dates of the following events, and give the names of the persons principally connected with them:

Martyrdom of Ridley.
Battle of Worcester.
Trial and execution of Strafford.
Assassination of Buckingham.
The completion of the present authorized version of the Bible.
Capture of Montrose.

3. What were the most important events in the reign of Queen Elizabeth?
4. Write a short life and character of Cranmer; and of Oliver Cromwell.

28 EXAMINATION PAPERS. [*Junior*

4. What charges were brought by their accusers against Mary, Queen of Scots, and the Earl of Strafford respectively? Where were they executed?
5. In what reigns did the following persons live, and in what way were they celebrated? Shakspeare, Simon de Montfort, Wycliffe, the Duke of Marlborough, Lord Bacon, Sir Isaac Newton, Lord Chatham.
6. State with regard to the following battles (1) the contending parties, (2) the victorious side, (3) the situation of the battle-field: Flodden, Cressy, Barnet, Culloden, Bosworth, Sedgemoor, Vittoria, the Boyne.

PART II.

TUESDAY, *Dec.* 14, 1858. 10 to 12.

II. 1. *a.* **Scripture.**

1. RELATE the chief offences of Saul, for which he was visited with God's anger.
2. Narrate the circumstances of the death of Saul and Jonathan. How did David receive the news? To whom did David say, "I will surely shew thee kindness for Jonathan thy father's sake"? How did he fulfil this promise?
3. Which of David's sons rebelled against him? How did these rebellions end?
4. Who were the Jebusites? Where was the seat of David's kingdom at first? To what place was it afterwards removed?
5. What was the first quarrel between the men of Israel and Judah?
6. Describe the numbering of the host and God's judgment on the people in consequence. In what did the sinfulness of David's act consist?
7. In what three ways was our Lord tempted in the wilderness?

Junior English History Paper, December 1858 (Cambridge University Archives Cam.c.II.51.18)

Greek History. This was expanded further to include Indian History by 1938, while the English History options were changed to two periods of British and European History. The Junior Examination was dropped in 1939 as a UK examination but remained as an overseas exam until 1953 during which time it was substantially revised.

Trends towards later periods of history caused fewer and poorer papers to be submitted on the early options, as highlighted in the 1939 examiners' report. Options in social history and American history began to emerge and with them came new comments and warnings from examiners: in 1945, for example, the disappointing results led to the advice that 'a candidate who does not know enough historical facts may be led to "waffle" on the social and economic questions.'

During this period a History Subject Committee emerged to manage the administration of History examinations and the development of the curriculum. The Committee was made up of History examiners, school teachers and senior officers from UCLES and recorded discord, transition and consensus in more or less equal measure. The early minutes show that although it held full discussions about syllabus criticisms, it was rather defensive and made little practical changes as a result. Criticisms were blamed on poor teaching and, in more than one case, on dislocation of schools after the upheavals of war. In contrast, specific requests for particular papers and questions by schools were met favourably, owing to 'book shortages in recent years' or those same upheavals of war. And so, for example, Irish History questions were introduced after a request received in October 1947.

In 1946 the School Certificate paper on the History of the British Empire was changed to History of the British Commonwealth and Empire. Here, too, there was an option on English, Social and Economic History. There was also a new special paper on West Indian History and the regulations draw attention to 'the provision of special History papers for other Oversea areas' which, it states, 'would be considered on application'.

In 1949, preparations for the new General Certificate in Education were finalised and it was decided that applications for specialist subjects would, in future, be refused. But the cultural shift towards greater variation had been made and the 1951 list of specialist subjects includes eight optional special subject papers, which were revised annually. As well as the new GCE O Level, the School Certificate became the new Oversea School Certificate for which there was a syllabus for the West Indies, the Sudan, Tropical Africa and Indian History. The GCE Examinations were not just new examinations but represented a new way of examining sixteen year olds. For the first time candidates could select a single subject without having to undertake a whole range of examinations as they had done in the past. From now on candidates would select History only if it was the right subject for them.

Question paper analysis

During this period we see the emergence of the classic pattern for School Certificate History of five essay questions in 2½ hours. We can be fairly confident that by this time individual questions were marked out of 20 with a paper total of 100.

The early question papers appear fairly similar to those at the end of 1917 although the papers at the end of the 1940s include several questions of a much more general nature. Rather than requiring candidates to describe a historical event or reign of a particular monarch, some questions focus on what life would have been like during a certain period, the increasing importance of music, what a typical parish church would have been like, and so on. There is a wider range of questions in the 1951 papers, including such general topics as the life of a colonist, pleasures and pastimes in town and country, the social and economic results of enclosures of the open fields, industry and agriculture.

The questions are the usual mix of describing and explaining though we can be reasonably sure that candidates and markers would not have perceived any essential difference. Point-by-point marking would award a mark for any relevant piece of information.

The English History paper is now clearly at least three papers within a paper, with ample question choice for teachers to cover no more than a single section – the rubric actually forbidding them from covering the entire period.

Most striking is the lack of pattern in the questions. There is, for example, no consistency of numbers of parts to each question or of command words. There is frequent use of either/or questions which serve only to increase the number of questions available. In today's terms this might constitute an assessment nightmare, but are very indicative of how the subject was perceived as a body of knowledge to be mastered, rather than as a set of skills to be acquired.

There exists an increasing expectation for candidates to hypothesise about the past and they would need to be able to produce third conditional structures or perfect conditional forms (e.g. 'would have done') in order to do so. There also appears to be a rising expectation that candidates would need to be able to produce a range of past tenses, active and passive verb forms, and be able to construct complex sentences and longer, coherently linked pieces of text.

The level of formality has been reduced with each paper: instructions are presented using imperatives and in the passive voice. Questions are constructed either using an imperative form or a question word. Past tenses, as well as present passive are commonly used. There is a continuing lack of consistency across the papers in, for example, use of articles and spelling. In one question 'organized' (with a 'z') appears. Candidates could be confused by inconsistencies in instructions such as 'describe in outline', 'outline' and 'describe briefly' and may feel that a different style of response is required for each. Interestingly, the pronouns 'she' and 'her' are used to describe countries. Countries or states within the British Empire are referred to as 'British possessions'.

There is a paper specifically for an overseas area – Indian History – though even in this paper there are questions on British History. Another new option is Modern European History (which in practice means the 19th century), indicative of the continuing trend away from English kings and queens.

Towards the end of this period as the School Certificate becomes, almost by default, the Oversea School Certificate, there is a new, more up-to-date paper on British and European History, but the question format shows no sign of change. This was just before the period of decolonisation which ushered in the processes of localisation of Cambridge Examinations, and brought about a whole range of History syllabuses for different countries and regions.

3. Post 1951

Overview

Of the 19,471 candidates who took GCE examinations in 1951, over 38% took an O level in History. The two papers on British and European History, 1688–1939 were by far the most popular. The other O level options were British and European History 1066–1714 and History of the British Empire and Commonwealth. School Certificate became the Oversea School Certificate in 1951 and the syllabus included all the options above plus papers on Indian History.

The minutes of the History Committee in October 1952 record that, 'The Examiners' Reports showed that the papers proved satisfactory to examiners and candidates' but there is no evidence from this source that after so much preparation and change the new examination settled into a rut. The committee discussed new options and ideas from schools and regions and during one meeting in October 1955 plans were put forward for a local history paper, an archaeology syllabus and a paper on Islamic History for West Africa. This, of course, was in addition to the annual revision of specialist subjects.

During this period the Syndicate was under pressure to examine later periods in history. A Committee of Secondary Teachers Association and the National Union of Teachers complained in 1968 that there were too few questions after 1918 and the Syndicate responded with the 'possibility of an additional paper which would cover twentieth century history'. The same report claimed that there were 'too many questions on wars and foreign policy' and so began a trend towards a History syllabus that is recognisable today.

The format of the examination was also reviewed during this period and the October 1968 Committee considered an alternative addition to the traditional essay type questions, 'proposing to experiment in the first place with a paper of short answer questions which can be objectively marked and which will provide a different kind of test to the one which is at present administered'. Also considered was 'a project scheme in which the teachers might make the first assessment of the work of their candidates'. An era of coursework had begun.

Despite progressive syllabus development, examiners, it seems, felt that candidates were not keeping pace. One examiner in 1969 claimed that 'many of the answers could have been written in the 1930s', while several others complained of narrow and out-of-date reading. The Report of 1972 covers familiar ground, warning candidates not to attempt questions covering too broad a period and to concentrate on answering the question. It also targets candidates' essay writing skills and 'poor organisation, leading to an ill-balanced arrangement of answers'.

The 1972 syllabus options are considerably more diverse than those for 1951. As well as the three British and European History O Level syllabuses and a syllabus on the History of the British Empire and Commonwealth, there were new or newish syllabuses on English Social and Economic History, World Affairs since 1919, and History of Europe, 1902–1964. Although entries had risen, the proportion of candidates for History had slumped to 13.5 % or 20, 786 entries with British and European History 1688–1939 still the most popular. For overseas candidates, however, History was still a popular subject: in Uganda and Kenya, only English Language and Geography attracted more candidates and, in Malaya, only Malay and Economics had higher entry figures. By 1971 the School Certificate syllabus included eleven options including specific papers on the History of India, Pakistan, Malaysia and Singapore and Central, Southern and East Africa.

Question paper analysis

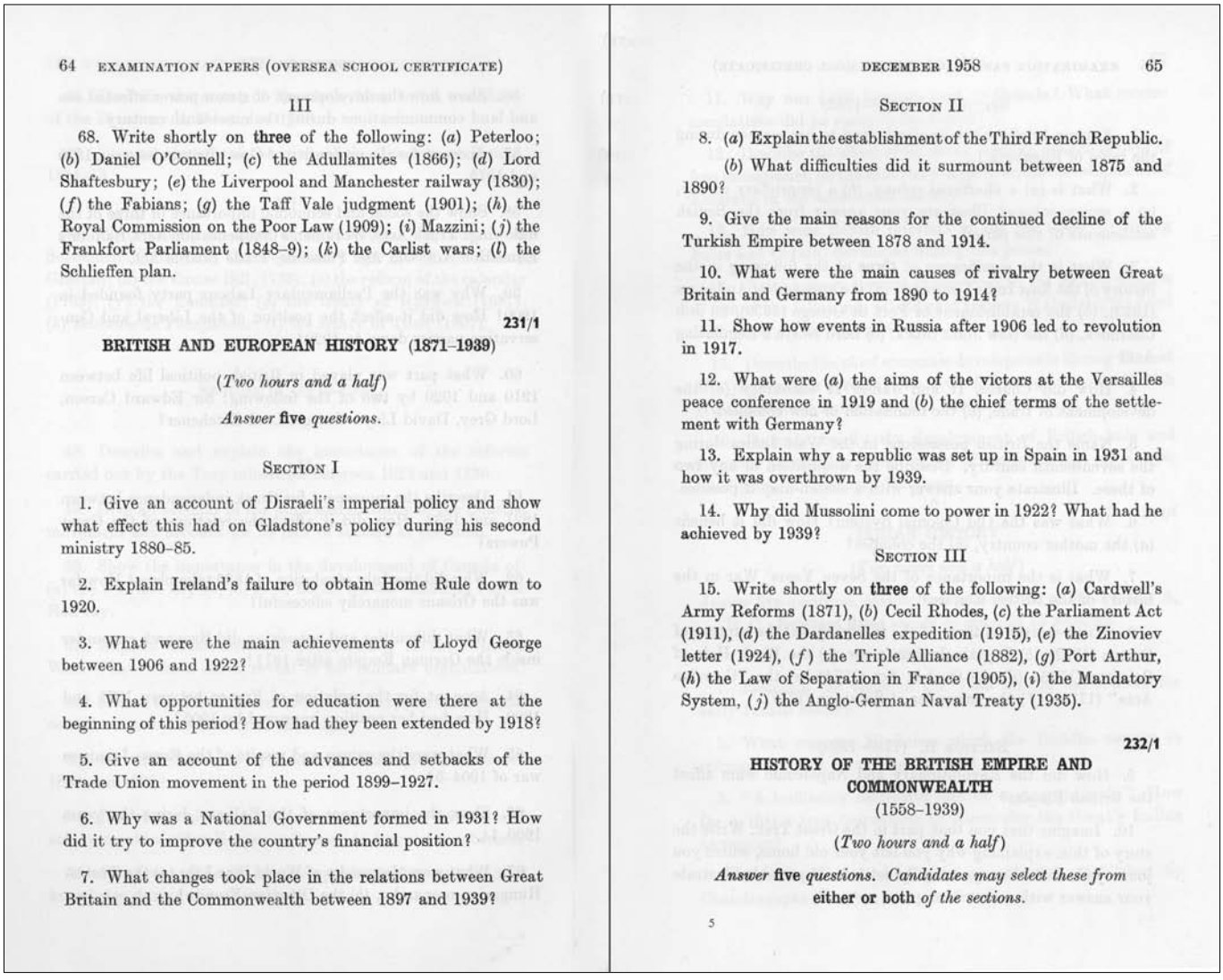
Apart from differences in course content, candidates in 1962 could have sat the 1934 question papers and seen nothing unfamiliar. The layout of the papers remains very similar during this period, with the same style and register of instructions and there is still no visual input or supporting text. One of the biggest changes in this period is the number of question papers available, and the wide range of topics included in the questions.

The thinking about curriculum change in History during this period focused almost entirely on what was to be taught rather than on how. As far as School Certificate was concerned, this meant new papers for different areas of the world, but the structure of these papers, and the nature of the questions on them, was almost always unvaried. The classic five essays in two and a half hours still held sway – all the more remarkable in that many candidates were not well equipped, particularly in their levels of English, for being tested in this manner.

The availability of papers in British and European History, first noted in 1940, continued with just minor date changes. They comprised enormous question choice so as to enable teachers to pick and choose whatever content they wished.

In the UK a new examination for those not able to take O Level, the Certificate of Secondary Education (CSE), was introduced in 1965. This gave examining boards the chance to explore new techniques for examining the less able. No such examination was available overseas, where candidates of all abilities were entered for the School Certificate (i.e. the same standard as O Level). This had implications for History which involved writing five essays, a demanding requirement for genuine O Level candidates, but perhaps impossible for those awarded School Certificate grades 7 or 8 (below O Level), or the even greater numbers who failed outright. Perhaps there was some recognition of this in the design of new syllabuses for African candidates where the assessment was split into two compulsory papers, each of one and a half hours, although candidates still had to answer three essay questions on each paper.

Overall the papers have a very similar feel to earlier ones both in style and linguistic terms. There is, however, increasing evidence of informality with instructions using imperatives and the second person, although the use of prepositions at the start of questions: 'Of what importance was China...', continue to indicate a more formal style. The active and passive voice is still used as is a range of verb tenses. Adverbs, with the old-fashioned collocation 'Write shortly' are still in evidence. However, more modern English is also in evidence, as the auxiliary 'did' has been used in a question with the verb 'have': 'What influence did West Indian planters... have on the British government...'



64 EXAMINATION PAPERS (OVERSEA SCHOOL CERTIFICATE)

III

68. Write shortly on **three** of the following: (*a*) Peterloo; (*b*) Daniel O'Connell; (*c*) the Adullamites (1866); (*d*) Lord Shaftesbury; (*e*) the Liverpool and Manchester railway (1830); (*f*) the Fabians; (*g*) the Taff Vale judgment (1901); (*h*) the Royal Commission on the Poor Law (1909); (*i*) Mazzini; (*j*) the Frankfort Parliament (1848–9); (*k*) the Carlist wars; (*l*) the Schlieffen plan.

231/1

BRITISH AND EUROPEAN HISTORY (1871–1939)

(*Two hours and a half*)

Answer **five** *questions.*

SECTION I

1. Give an account of Disraeli's imperial policy and show what effect this had on Gladstone's policy during his second ministry 1880–85.

2. Explain Ireland's failure to obtain Home Rule down to 1920.

3. What were the main achievements of Lloyd George between 1906 and 1922?

4. What opportunities for education were there at the beginning of this period? How had they been extended by 1918?

5. Give an account of the advances and setbacks of the Trade Union movement in the period 1899–1927.

6. Why was a National Government formed in 1931? How did it try to improve the country's financial position?

7. What changes took place in the relations between Great Britain and the Commonwealth between 1897 and 1939?

DECEMBER 1958 65

SECTION II

8. (*a*) Explain the establishment of the Third French Republic.

(*b*) What difficulties did it surmount between 1875 and 1890?

9. Give the main reasons for the continued decline of the Turkish Empire between 1878 and 1914.

10. What were the main causes of rivalry between Great Britain and Germany from 1890 to 1914?

11. Show how events in Russia after 1906 led to revolution in 1917.

12. What were (*a*) the aims of the victors at the Versailles peace conference in 1919 and (*b*) the chief terms of the settlement with Germany?

13. Explain why a republic was set up in Spain in 1931 and how it was overthrown by 1939.

14. Why did Mussolini come to power in 1922? What had he achieved by 1939?

SECTION III

15. Write shortly on **three** of the following: (*a*) Cardwell's Army Reforms (1871), (*b*) Cecil Rhodes, (*c*) the Parliament Act (1911), (*d*) the Dardanelles expedition (1915), (*e*) the Zinoviev letter (1924), (*f*) the Triple Alliance (1882), (*g*) Port Arthur, (*h*) the Law of Separation in France (1905), (*i*) the Mandatory System, (*j*) the Anglo-German Naval Treaty (1935).

232/1

HISTORY OF THE BRITISH EMPIRE AND COMMONWEALTH

(1558–1939)

(*Two hours and a half*)

***Answer* five** *questions. Candidates may select these from* **either or both** *of the sections.*

5

Oversea School Certificate History Paper, December 1958 (Bound Volume, 1958)

4. IGCSE

Overview

During this period the move towards more modern history gathered momentum and became more pronounced. By 2000 the core curriculum for the GCSE included no papers on pre-twentieth century history. Other new elements included the 'in depth' part of the syllabus, and the coursework. The coursework in the 1989 paper carried 30% of the marks. But this trend was reversed during the period and by 2000 the coursework element had dropped to 25%.

Historiography, the discussion and analysis of original source material, had become a feature of the examinations, but was revised in 1987 on the advice of teachers concerned that the language in primary source material used for GCE O Level was not suitable for use at GCSE.

The 1989 and 2000 GCSE History syllabuses were made up of a compulsory core and optional papers. The 1989 syllabus options all comprise two papers and coursework, with source-based questions included in all papers. The syllabuses for 1989 continued to follow periods in history in all cases except the School History Project. By 2000 however, theme-based study had filtered into all the History options which had been rationalised to just three: the Schools History Project, Modern World and British & Social Economic History. But the new type of syllabus does not lack diversity. It includes a range of thematic studies from which candidates could choose, such as, Medicine through Time and Germany, 1919–1945.

These changes were made through a substantial consultation process. In 1987 the History Subject Committee asked that Examiners' Reports include 'additional guidance to schools on how most effectively to prepare candidates for examinations', a sharp contrast to the attitude towards teachers in the 1940s. A Consultation Document for MEG GCSE History Syllabus 1990 proposed amendments to all but the School History Project syllabus, commenting 'it is felt that the revisions to the core will introduce an element of flexibility and choice which will amount to a significant reduction in the content burden faced by candidates.' It included the abolition of short answer questions, introduced in the 1960s, which perhaps reflects the crisis felt at the falling number of A Level History candidates during this period.

As the overseas examination, the IGCSE followed the same pattern of a compulsory core and, in this case, a regional optional paper. The preference for Modern History was more pronounced in the IGCSE right from the start, focusing on periods after 1919. The regional options included papers on Southern Africa, the Americas, Western Europe, USSR & Eastern Europe, Eastern Asia or Middle East & Eastern Mediterranean. Additionally, candidates were obliged to submit a school-based assessment or take an alternative to coursework paper – a topic from the core content which involved a series of questions on a collection of source material.

The June 1989 report on IGCSE shows that the new examination was doing well. Entry figures had doubled from the previous year, bringing in 'more centres where candidates had properly studied the course and understood what they were attempting'. Examiners, too, praised the 'surprisingly high' quality of writing. In 2000 the IGCSE also underwent revision to break up the elements of factual material and the use of sources, which had previously existed together in one paper. The focus for the core content and depth studies remained with the twentieth century and the response to these changes was greeted positively in the examiners' reports.

A cartoon printed in the 'Daily Express' in 1943.

(a) Look carefully at the cartoon.
(i) Why is W. Beveridge described as the 'architect' of Social Security? [1]
(ii) Explain the term 'Social Security' as used in the cartoon. [3]

(b) When the war ended in 1945, why was the idea of a 'welfare state' so popular amongst the British people? [6]

(c) How successful were the Labour Governments of 1945–51 in creating a 'welfare state'? [15]

IGCSE History Paper 0470/4, June 1989 (1989/2/2)

Question paper analysis

IGCSE gave an opportunity to consolidate curriculum developments of the previous decade and move into the mainstream for candidates of all levels of ability. For History, this meant the adoption of ideas pioneered by the Schools Council History Project, which stressed a skills-based approach to the subject and to History examinations.

IGCSE papers are quite different to those seen in previous years particularly with the extensive use of visuals and supporting text. The use of high-level input information to set the scene for questions suggests that emphasis is being placed on the top-down processing model of language or reading comprehension. This is a model based on the belief that readers make sense of discourse by moving from the highest units of analysis to the lowest, and that comprehension is achieved by firstly activating background knowledge or schemata and setting the context. There are plenty of examples of structurally complex input including: cleft sentences – 'It was the election of Lincoln as President that made war certain'; organisation in terms of desired thematic prominence rather than for accessibility or simplicity of structure; and reported speech using a range of verb tenses.

There is now a markedly different layout to earlier papers. The main difference is visuals in the form of photographs, maps, graphs and other illustrations which are included with many of the questions. There are quotes from speeches, extracts from books and newspapers and statistics, all used to set a context or give support to questions which follow. This means that questions are much longer than they have been previously, some taking up a page of space.

There is also metaphoric use of language in some questions reflecting the radical change in approach to history study and teaching at all levels and ages, that is, a move away from the recall of facts and study of definitive works to a more historiographical approach.

Optional questions are all set to a standard pattern, marks available are printed on the question paper, essay questions are structured into three parts to help the less able, and stimulus material is used, again as an aide-memoire for the less able.

Underpinning all this is an explicit statement of assessment objectives in the syllabus document. Everyone knows what is being tested and where. In the structured essays, for example, part (a) tests recall, part (b) tests understanding of causation, and part (c) tests the ability to construct an argument. This is considered crucial to the study of History today. Perhaps most radical is the inclusion of a section of the paper testing skills of handling historical sources.

The demands made on candidates by these papers are very different from those on the old School Certificate papers. To reflect this, the method of marking was also changed so that marks would be awarded according to the quality of explanation, or the level of skill shown in the answer, and not because of the amount of factual knowledge demonstrated. It is now appreciated that giving less able candidates materials with which to work, like a collection of sources, helps to provide them with a basis for their answers. The most difficult questions are those which give them no such structure, such as essays.

No fundamental change in the IGCSE History examination had occurred by 2000, though the papers had been slightly reformatted. Source evaluation was given a paper in its own right, and the structured essays were all consolidated into a single paper. The options within the syllabus had been increased slightly, most notably to offer a 19th Century path through a syllabus which formerly had been exclusively drawn from 20th Century World History.

Although the input material in these IGCSE papers is significantly more complex and of a higher level, the instructions and rubrics are much clearer and more accessible and there is evidence that the rubrics and instructions in the later paper have been simplified further, so the questions themselves are very clear and easy to understand. Despite the lexical input being of a higher level than seen previously, the output would not necessarily need to be different. Candidates would need to produce a wide range of lexis throughout.

The expected output is made much clearer by indicating the total number of marks available for each section: this would enable candidates to judge more effectively how much time and effort to invest in each part.

Conclusions

There are huge differences between the earliest 1858 question papers and those from 2000, in terms of length, topics tested, presentation, level of formality, and linguistics. Looking at the papers in the intervening years, these differences appear gradually, with the most dramatic change taking place between 1972 and 1989. The inclusion of visuals and supporting text from 1988 means not only that it is considered important to set the context and activate candidates' background knowledge before focusing on specific details, but also that the level of linguistic input is much higher than previously seen.

The changes in the papers over the years reflect the style of teaching methodology that was popular at that time:

- from rote-learning in the mid-19th century,
- to a focus on interpretation and opinion in the early to mid 20th century,
- and the belief that discourse is interpreted using top-down processing strategies in the late 1970s/1980s.

Linguistically, the biggest change is in terms of the complexity of language used in the stimulus material. Although the lexical level of questions is high throughout the years, the last two sets of question papers are undoubtedly more complex and candidates would need a higher level of comprehension in order to cope with some of the authentic extracts from speeches or printed texts. Conversely, the level of lexical and structural input in the instructions has been steadily simplified and made clearer. Although the level of linguistic input has definitely changed, the expected level of output seems to be fairly constant.

From the papers selected for inclusion in this study one would conclude that the nature of the Syndicate's History examinations changed surprisingly little in the century after 1858. For the earliest papers it is now hard to infer accurately what the marking processes were but by the 1920s a model of testing History had been established that then lasted, essentially unchanged, for more than fifty years. Whilst it would be prudent to exercise some caution about the idea that IGCSE changed everything overnight, at no other time since 1858 has the nature of History as a school subject been so fundamentally rethought, with consequent changes in the processes of its assessment.

Sources from Cambridge Assessment Archives (unless otherwise stated)

Comments by Subject Committees on criticisms of question papers received through the Standing Joint Committee for Cambridge Examinations of the Joint Four Secondary Teachers' Associations and the National Union of Teachers, March 1969. Bound Volume, 1968

Examination Set Texts for Colonial Centres, 1893–1914. Archive Ref: A/ST 1/15

Examiners' Report from Bound Volumes, 1939, 1972

Examiners' Report (IGCSE), June 1989; 1989/4/2

Examiners' Report (IGCSE), June 2000; 2000/4/2

History Subject Committee Papers, June 1935 – Feb 1957 Archive Ref: S/H 2/1

History Subject Committee Papers, Feb 1945 – Oct 1969 Archive Ref: S/H 2/2

History Subject Committee Agenda Papers, 22nd Oct 1987 Archive Ref: S/H 1/16

J N Keynes Commonplace Book, 1882 – 1910, Extracts, 1883/4. Archive Ref: PP/JNK 1/3

Oversea Awarding Committee Minutes, 1959 – 1972 (1971). Archive Ref: C/OAC 2/1

Pass lists from Bound Volume, 1917

Question Papers, Junior, 1858 Originals at Cambridge University Library. University Archives Ref: Cam.c.II.51.1

Question Papers from Bound Volumes (Juniors), 1884,1909, 1917, 1928

Question Papers from Bound Volumes (School Cert.), 1928, 1934, 1940, 1951, 1962, 1972

Question Papers from Bound Volumes (GCE O level), 1962, 1972

Question Papers (IGCSE), 1989; 1989/2/2

Question Papers (IGCSE), 2000; 2000/2/2

Regulations from Bound Volumes (Junior Examinations): 1874, 1890, 1899, 1951

Regulations from Bound Volumes (all examinations): 1917, 1928

Regulations from Bound Volumes (School Certificate): 1938, 1939, 1946, 1951, 1971

Regulations and Syllabus from Bound Volume (School Certificate), 1971

Regulations and Syllabus from Bound Volume (GCE O level), 1972

Syllabus (IGCSE), 1989; 1989/1/2

Syllabus (GCSE), 1989; 1989/1/1

Syllabus (GCSE), 2000; 2000/1/1

Syllabus (IGCSE), 2000; 2000/1/2

UCLES Annual Reports (Examiners' Reports and Statistics): 8th (1866), 29th (1877), 42nd (1899), 45th(1902), 56th (1913), 58th (1915), 93rd (1951)

The reliabilities of three potential methods of capturing expert judgement in determining grade boundaries

Nadežda Novaković and Irenka Suto Research Division

Introduction

In England there is a strong public expectation that qualification standards should remain constant over time. For example, a candidate who achieves a grade B in GCSE Spanish in one year should be considered 'comparable' in some sense to candidates from previous years who also achieved a grade B in GCSE Spanish. At each examination session, awarding bodies must therefore determine the grade boundaries for their examinations that equate to those of previous sessions. A great deal of research activity is directed towards investigating different methods for capturing the expert judgement of professionals who are given the responsibility of determining grade boundaries and thus maintaining year-on-year examination standards.

In this article, we report the results of some research[1] investigating the reliabilities of three such (potential) methods for capturing expert judgement, as used in:

(i) Traditional (current) awarding

(ii) Thurstone pairs

(iii) Rank ordering.

The traditional awarding method is the principal method used operationally for determining grade boundaries in the context of public examinations and England, while rank ordering and Thurstone pairs have been sometimes suggested as alternatives to the judgemental process used in traditional awarding.

Traditional awarding

When the traditional awarding method is used, a committee of senior examiners (led by a Chair of Examiners) looks at a sample of candidates' scripts in the mark range where the grade boundary is expected to be. They are required to make holistic, absolute judgements about whether each script on a particular mark is worthy of the grade in question, for example, 'this script is worthy of grade A' or 'this script is a borderline grade B script'. This type of judgement implies that examiners (judges) have an internal standard about what, for instance, a grade A script should look like; it is assumed that judges would have internalised this standard partly through experience and partly through studying archive scripts.

The judges decide on the lowest mark for which there is consensus that the work is worthy of the higher grade and the highest mark for which there is consensus that the work is not worthy of the higher grade. This gives a range of marks called the 'zone of uncertainty', or simply 'zone'. The committee then use their collective professional judgement, referring to statistical information on the overall performance of the examination, to recommend an appropriate grade boundary from within that range. Throughout the process, judges have access to 'archive' scripts from the previous year's examination, with marks on the equivalent grade boundary. Statistical information on performance on individual questions may also be available.

Concerns have been raised over the reliability of the judgements made in the traditional awarding method (Willmott & Nuttall, 1975; Greatorex & Nádas, 2008). Good and Cresswell (1988) replicated some awarding meetings for GCSE French, History and Physics and found that parallel groups of judges reached slightly different decisions about grade boundaries, which, if substituted for one another, would have affected the grade of 13% of French candidates, 17% of physics candidates and 38% of history candidates. Imperfect reliability may stem from the method's reliance on absolute judgements. Drawing on Laming's theory of the nature of human judgement (2004), that absolute judgement cannot occur and that all judgements are comparisons of one thing with another, Raikes *et al*. (2008) have argued for replacing traditional awarding with methods in which judges make relative judgements about the quality of candidates' work.

A recent empirical study by Gill and Bramley (2008) supports this view. The study's participants were experienced history and physics examiners who were given pairs of scripts and asked to make absolute judgements about the grade each script deserved. The participants also made relative judgements about the pairs of scripts, that is, they judged which of the two scripts was better in terms of overall quality. All scripts were cleaned of marks and the participants had no reference to archive scripts or any statistical information. The examiners' judgements were compared with the marks and grades that the scripts originally received, and the results showed that examiners had difficulty in replicating the decisions made at the live awarding meetings which they themselves had attended: the percentage of judgements matching the original grades was below 40% for history and below 25% for physics. On the other hand, the overall accuracy of the relative judgements was higher than that of the absolute judgements (history examiners ordered 66% of the paired comparisons in correct mark order, while physics examiners ordered 78% of the comparisons in correct mark order).

Methods using relative judgements

In view of the criticisms levelled against the traditional awarding method, Thurstone pairs and rank ordering have been suggested as possible replacement methods of capturing expert judgement in determining

1 The wider research project also addressed an aspect of the validities of these methods by investigating and comparing the features of candidates' work that most influence experts in each method; these results were presented by Novaković and Suto (2009).

grade boundaries. Both methods rely on examiners making relative holistic judgements about the quality of candidates' work, which arguably have more psychological validity than absolute judgements. Furthermore, judgements made in rank ordering and Thurstone pairs are not influenced by statistics or by candidates' marks, which are always visible in traditional awarding (see Black & Bramley, 2008, and Greatorex, 2007 for a detailed list of advantages of rank ordering and Thurstone pairs over the traditional awarding method).

Thurstone pairs

In recent decades, the Thurstone pairs method (Thurstone, 1927a, b) has been used in comparability studies in the UK and internationally. In this method, judges are required to individually compare pairs of candidates' scripts from two different examinations (for example, from two different years). For each of many pairs of scripts, the judge must decide which candidate's performance is better (no ties are allowed). The scripts are often cleaned of marks, which are on or near the grade boundary under consideration. If these comparisons are repeated many times, then Rasch analysis can be used to place all scripts from both examinations on a single common scale of measurement, representing a latent construct of script quality. The equivalent marks of the different examinations can then be calculated, enabling standards to be compared (see Bramley, 2007).

Kimbell *et al*. (2007) are the first to have investigated the use of Thurstone pairs as a method for harnessing expert judgement in grading, but no systematic comparisons with the outcomes of more conventional methods of grading have been carried out. Hence, there are no established procedures for using Thurstone pairs in grading. The main drawback of the Thurstone pairs method is that it can be time consuming, particularly when considering a large number of scripts, which take time to read and might be remembered, thus probably violating the requirement that each paired comparison should be independent of any previous comparison.

Rank-ordering

The rank ordering method (Bramley, 2005) is similar to Thurstone pairs in that judges individually compare candidates' scripts (which have been cleaned of marks) from two different examinations. However, rather than judging which of a pair of scripts is better, the judge must rank individual scripts in a pack, in order of overall quality. Half the scripts in the pack are from one examination and the other half are from the other examination. Judges repeat the process with a number of packs of scripts, and scripts from the whole range of marks are used. Each judge has a different combination of scripts in their packs. As with Thurstone pairs, Rasch analysis enables all scripts from both examinations to be placed on a single scale of measurement; the equivalent marks (and grade boundaries) can then be calculated. Rank ordering is more time-efficient than Thurstone pairs and it can be designed to ensure that the number of times a judge sees a particular script is minimised, reducing the possibility of the scripts being remembered.

The rank-ordering method has been used for the purposes of setting grade boundaries, both in an operational setting (for Key Stage 3 English examination, see Bramley, 2005) and in research settings (Black, 2008; Black & Bramley, 2008; Elliott *et al*., 2005; Gill & Black, 2006).

Black and Bramley (2008), and Gill *et al*. (2007) have investigated whether traditional awarding and rank ordering generate the same grade boundaries, by using these two methods to cross-validate the traditional awarding of an A-level psychology paper and GCSE English paper respectively. Both studies found some concurrence and some disparity at key grade boundaries. However, given that traditional awarding uses a blend of both judgemental and statistical information, the methods' outcomes should not be expected to be identical.

An adaptation of the rank ordering method has recently been used experimentally by Raikes *et al*. (2008) in the context of an AS-level biology examination. Research participants were required to judge the relative qualities of sets of three scripts at a time. Four groups of judges were involved in the study: members of the existing awarding committee; other examiners who had marked the scripts operationally; teachers who had taught candidates for the examinations but not marked them; and university lecturers who teach biology to first year undergraduates. Raikes *et al*. identified very high levels of intra-group and inter-group reliability for the scales and measures estimated from all four groups' judgements.

The present study

We conducted a three-way comparison of the intra-method and inter-method reliabilities of all three methods in the context of setting grade boundaries.

Intra-method reliability refers to the comparison of the grade boundaries yielded by each single method in turn, if used by different groups of judges and on different sets of scripts. While the literature indicates that the intra-method reliability of traditional awarding is imperfect, it is unclear how it compares with that of the Thurstone pairs and rank ordering methods when these are used in grading. To our knowledge, a direct comparison has not previously been made.

Inter-method reliability refers to the comparison of the grade boundaries that the three methods would yield if used on the same examination papers. Arguably, high inter-method reliability would suggest that judgements are made in reference to a common construct (or a common subset of constructs). The above-mentioned studies by Black and Bramley (2008), and Gill *et al*. (2007) have addressed this issue to some extent by comparing the outcomes of the traditional and rank ordering methods. However, this issue is clearly ripe for further investigation. The Thurstone pairs method has not been compared directly with either of the other two methods in the context of standard maintenance.

Experimental design

The research focused on two written examination papers with contrasting question and response styles, which were administered by OCR examinations in June 2007 (available from www.ocr.org.uk). One paper (maximum mark = 45) was from an AS-level biology syllabus, and the other paper (maximum mark = 90) was from a GCSE English syllabus. The research was carried out using samples of past candidates' scripts: for biology, the research focussed on the E/U and A/B grade boundaries; for English, the research focussed on the C/D and A/B boundaries.

The experimental design was identical for biology and English, taking the form of a 3×3 'Latin square' (see Table 1). For each subject, three mutually exclusive sets of examination scripts were created, which were matched for mark. Three groups of ten 'judges' (examiners, matched for experience of the methods) made judgements using each of the three methods on a different set of scripts. Thus, each judge group encountered the three methods in a unique order, and ultimately, judgements of each method were conducted on all three script sets. The Latin square design

thereby enabled comprehensive comparisons of the three methods, whilst controlling for order effects.

Table 1: Latin square design

Judge group	*Script set and order of attempting tasks*		
	1	*2*	*3*
1	Rank ordering	Traditional awarding	Thurstone pairs
2	Thurstone pairs	Rank ordering	Traditional awarding
3	Traditional awarding	Thurstone pairs	Rank ordering

Procedure

Each judge received three sets of photocopied scripts (one for each of the tasks) together with a covering letter, detailed instruction sheets for individual tasks, statistical information on the candidates for use in the traditional awarding task, charts for recording judgements, and copies of the question papers and mark schemes from June 2007 and June 2006.[2] The judges were given three weeks to complete the tasks from home and were advised to take about half a day per task. They were asked to (re)familiarise themselves with the question papers and the mark schemes before embarking on the tasks. Judges were asked not to re-mark the scripts; instead, they should make a holistic judgement about each script's quality.

For each task, each group of judges used scripts drawn from a different script set (see Table 1). Within each script set, the numbers and marks of scripts selected for use in each judgemental method were determined by the common practice for that method. (Script selection for Thurstone pairs followed previous studies (Bell *et al*.,1998; Bramley *et al*.,1998)).

Traditional awarding

Biology judges received ten scripts around the E/U boundary and ten scripts around the A/B grade boundary. They also received four 'archive' scripts from June 2006 – two on each grade boundary mark. English judges received twelve scripts around the C/D boundary and twelve around the A/B boundary, as well as four 'archive' scripts – two on each grade boundary mark. The judges' task was to decide whether the June 2007 scripts were worthy of the grade under consideration. The scripts' marks were clearly visible.

Thurstone pairs

For each subject, the judges received two packs of scripts. Pack 1 contained a total of 20 scripts around the higher boundary, while Pack 2 contained a total of 20 scripts around the lower boundary. In each pack, 10 scripts were from June 2006 and 10 scripts from June 2007. The judges compared two scripts at a time, and judged which represented the better performance.

Rank ordering

For each subject, the judges received four packs of scripts. Each pack comprised 10 scripts: 5 from 2007 and 5 from 2006. Each pack contained a unique selection of scripts, but there were common scripts between the judges' packs allowing each entire set of scripts to be linked. The task included all the scripts that were used in the study and these covered the entire mark range for both examinations. The judges ranked the scripts in each pack in the order of their relative quality.

Analysis of grade boundary data

All judges completed the tasks successfully. The analytical methods for determining grade boundaries were different for traditional awarding on the one hand, and for Thurstone pairs and rank ordering methods on the other. All judgements from the traditional awarding task were sent to the appropriate Chairs of Examiners, who were asked to look at the judges' decisions and determine the zones of uncertainty and grade boundaries for each judge group.

For the rank ordering data, FACETS software (Linacre, 2005) was used to employ multi-faceted Rasch analysis, which allowed scripts from 2006 and 2007 to be placed on the same scale of perceived quality. The raw mark scales of the two examinations could then be compared directly so that mark x in one year could be deemed equivalent to mark y in the other year in terms of perceived quality of candidate performance.

For Thurstone pairs, Rasch analysis was also employed. However, due to the very restricted mark ranges of the scripts used, (which were very close to the grade boundaries), it was inappropriate to directly relate the mark scale to the scale of perceived quality in this case. We therefore used a crude method of calculating the equivalent marks, which used the following formula:

2007 Thurstone implied boundary =

2007 mean mark - [(SD 2007 mark/SD 2007 measure)

× (Mean 2007 measure – Mean 2006 measure)].

The boundary marks generated by the Thurstone pairs task therefore have to be viewed with some caution.

Findings relating to grade boundaries

The grade boundary marks for 2007 that were generated experimentally by the three methods are summarised in Tables 2 and 3 (biology), and 4 and 5 (English).

For biology, intra-method reliability was excellent for traditional awarding: the boundary marks generated were identical across the three judge groups for one boundary, and identical for two judge groups on the other boundary. The reliability of Thurstone pairs was also very high: for both grade boundaries, the boundary marks were identical for two judge groups, while the boundary mark of the third group differed by only one mark. The intra-method reliability of rank ordering was slightly lower but still very high: it was perfect for the A/B grade boundary, but for the E/U boundary three different boundary marks were generated, all one mark apart.

For English, the findings were similar. Although for four of the six boundaries to be determined, the Chair of Examiners felt unable to complete the task without referring to statistical indicators, the zones of uncertainty restricted potential grade boundaries to such an extent that it was still possible to conclude that the intra-method reliability of traditional awarding was high. The intra-method reliability of the

2 In a linked study, the judges also completed a fourth task in which they rated scripts on a number of different features. This was part of a wider research project, presented by Novaković and Suto (2009).

Table 2: Summary of E/U grade boundary marks for biology

Task	*Judge group*			*Actual 2007 grade boundary mark*
	Group 1	*Group 2*	*Group 3*	
Traditional awarding	16	16	16	17
Thurstone pairs	15	16	15	
Rank ordering	14	14	14	

Table 3: Summary of A/B grade boundary marks for biology

Task	*Judge group*			*Actual 2007 grade boundary mark*
	Group 1	*Group 2*	*Group 3*	
Traditional awarding	35	34	34	34
Thurstone pairs	33	33	32	
Rank ordering	32	31	33	

Table 4: Summary of C/D grade boundary marks for English

Task	*Judge group*			*Actual 2007 grade boundary mark*
	Group 1	*Group 2*	*Group 3*	
Traditional awarding	56	55	? (54–56)	55
Thurstone pairs	55	55	56	
Rank ordering	56	57	58	

Table 5: Summary of A/B grade boundary marks for English

Task	*Judge group*			*Actual 2007 grade boundary mark*
	Group 1	*Group 2*	*Group 3*	
Traditional awarding	? (69–70)	? (69–70)	? (68–70)	69
Thurstone pairs	69	70	69	
Rank ordering	69	68	72	

Thurstone pairs method was also very high. For both grade boundaries, two groups generated the same boundary mark, whereas the mark of the third group differed by only one mark. Intra-method reliability was again lower for rank ordering. For the C/D grade boundary, three different boundary marks were generated, all one mark apart. For the A/B grade boundary, all three boundary marks were different, and spanned a five-mark range.

There was no overall trend in leniency/severity across the judge groups for either subject: no single group generated boundary marks that were consistently higher or lower than the marks of the other two groups. This finding may be taken to confirm that the judge groups in the study were well matched.

When the three methods are compared with one another, it appears that for both subjects, the traditional awarding and Thurstone pairs methods generated very similar boundary marks, except for the biology A/B grade boundary. The boundary marks generated by rank ordering were all on the lenient side for biology, whereas for the English C/D grade boundary, they were on the severe side. However, without using additional research methods to triangulate findings, it is not possible to determine which of these June 2007 grade boundary marks are equivalent ontologically to the actual 2006 boundary marks. It is therefore not possible to conclude from this study which method, if any, is ultimately the most effective at maintaining standards.

Limitations

While the research has found traditional awarding to have high intra-method reliability, there is a possibility that this reliability is simply an artefact of the method – even if the 'zone' had been as wide as the mark range, it is possible that the boundary mark would still have been chosen in the middle. A possible way of investigating intra-method reliability of traditional awarding in more detail would be to give different groups of examiners scripts covering non-identical mark ranges (offset by a few marks) and ask them to set the grade boundaries. In our study, however, we wanted to keep the procedure as close as possible to the one used at live awarding meetings.

One of the major limitations relates to the way that the Thurstone pairs method was used in our study, that is, for the purpose of producing grade boundaries. As there is no existing procedure for using Thurstone pairs as a grading method, we used it as it has been used in comparability studies, using scripts only in a small range around the grade boundary. This made it impossible to calculate equivalent marks by plotting pairs of regression lines (as in rank-ordering), and the grade boundary marks for this method were calculated using an alternative and rather crude method. These marks therefore need to be regarded with caution.
A better way of using Thurstone pairs for grading purposes would be to use the scripts covering a wide mark range, although this might prove impractical or tiring, considering the number of judgements that would need to be made. Kimbell *et al*. (2007) have been using Thurstone pairs for grading purposes on a wide range of marks; however, they have used Thurstone pairs in combination with rank-ordering (thus creating a hybrid grading method), and they have so far not proposed a way of translating the experts' judgements into the actual grades.

A limitation of all three methods is their reliance on particular individuals for critical judgements. For traditional awarding, the zones of uncertainty and grade boundaries were judged by Chairs of Examiners alone, as it was impractical for them to harness the other judges' collective professional judgement. For Thurstone pairs and rank ordering, the researchers made equally crucial judgements during the Rasch analyses, about which misfitting or outlying scripts and judgements to exclude.

Conclusions

It can be concluded from this study that, reassuringly, none of the three methods investigated is strikingly weak in terms of either type of reliability, and all three methods appear to have functioned well, generating highly plausible grade boundaries. Whilst theoretically, methods that rely on comparative rather than absolute judgements might be favourable (Laming, 2004), this study provides no empirical evidence to support such a preference. The implication of this is that any of the three methods explored could contribute to the determination of grade boundaries operationally.

Table 6: Final comparison of traditional awarding, Thurstone pairs and rank ordering

	Traditional awarding		*Thurstone pairs*		*Rank ordering*	
	Biology	*English*	*Biology*	*English*	*Biology*	*English*
Intra-method reliability	Excellent	Very high	Very high	Very high	Very high	Reasonable
Inter-method reliability	Quite high with Thurstone pairs and rank ordering	Very high with Thurstone pairs; quite high with rank ordering	Quite high with traditional awarding and rank ordering	Very high with traditional awarding; quite high with rank ordering	Quite high with traditional awarding and Thurstone pairs	
Number of judgements made per judge in ½ a day's work	20	24	40	40	40	40
Key operational advantages	No need for scripts to be cleaned of marks.		Requires no extra input from the Chair of Examiners.		Requires no extra input from the Chair of Examiners.	
Key operational disadvantages	Requires considerable input from Chair of Examiners.		Scripts must be cleaned of marks (less problematic for scripts marked on-screen). Requires a large quantity of archive scripts.		Scripts must be cleaned of marks (less problematic for scripts marked on-screen). Requires a large quantity of archive scripts.	
Key theoretical strengths	Draws on the collective expertise of 'communities of practice', though only while meetings continue to be largely face to face. Arguably, remote awarding risks weakening these communities.		Relies on relative rather than absolute judgements. Unaffected by judges' leniency or severity.		Relies on relative rather than absolute judgements. Unaffected by judges' leniency or severity. Large number of paired comparisons obtained from actual human judgements.	
Key theoretical weaknesses	Relies on absolute rather than relative judgements. Affected by judges' leniency or severity. Judgements are 'contaminated' by statistical information.		Rasch techniques (e.g. FACETS) are often used to analyse the data – the modelling assumption of a single latent trait is controversial. The mark range covered by scripts is too small to calculate equivalent marks without making considerable assumptions. When scripts cover a wide mark range, the judges' task can become tiresome, and rank-ordering lends itself better to producing a large number of comparisons.		Rasch techniques (e.g. FACETS) are often used to analyse the data – the modelling assumption of a single latent trait is controversial. Places significant demands on the working memory.	

Overall, the results of our study do not provide enough evidence to favour one method over the other two, either for operational or research purposes. However, in Table 6 we have drawn together the findings from our study and from other research and anecdotal evidence relating to the three methods. We hope it will prove useful to anyone making a decision about which method to use. It is important to emphasise once again that while rank ordering has been used for grading purposes previously, there is no existing procedure for using Thurstone pairs in determining grade boundaries. In this table, we have listed the advantages and disadvantages of Thurstone pairs as it has been used in this study (adapted from comparability studies).

References

Bell, J.F., Bramley, T. & Raikes, N. (1998). Investigating A level mathematics standards over time. *British Journal of Curriculum and Assessment* **8**, 2, 7–11.

Black, B. (2008). *Using an adapted rank-ordering method to investigate January versus June awarding standards*. Paper presented at the Fourth Biennial EARLI/Northumbria Assessment Conference, 27–29 August in Berlin, Germany.

Black, B. & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education* **23**, 3, 357–373.

Bramley, T. (2005). A rank ordering method for equating tests by expert judgment. *Journal of Applied Measurement* **6**, 2, 202–223.

Bramley, T. (2007). Paired Comparison Methods. In: P. Newton, J. Baird, H. Goldstein, H. Patrick and P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Bramley, T., Bell, J.F. & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone Paired Comparisons. *Educational Research and Perspectives* **25**, 2, 1–23.

Elliott, G., Johnson, N. & Bramley, T. (2005). *Cross-validation of 2004 standard setting in GCE AL Psychology 2540 using a rank-ordering methodology*. Cambridge Assessment internal report.

Gill, T. & Black, B. (2006). *An investigation of standard maintaining and equating using expert judgment in GCSE English between years and across tiers using a rank-ordering method*. Cambridge Assessment internal report.

Gill, T., Bramley, T. & Black, B. (2007). *An investigation of standard maintaining in GCSE English using a rank-ordering method*. Paper presented at the British Educational Research Association Conference, 5–8 September in London, UK.

Gill, T. & Bramley, T. (2008). *How accurate are examiners' judgments of script quality?* Paper presented at the British Educational Research Association Annual Conference, 3–6 September in Edinburgh UK.

Good, F.J. & Cresswell, M.J. (1988). *Grading the GCSE*. London: Secondary schools Examination Council. Referenced in: M. Cresswell. (2000), *Defining, Setting and Maintaining Standards in Curriculum-Embedded Examinations: Judgemental and Statistical Approaches*. In: H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, developments and statistical issues*, 57–84. Chichester: John Wiley and Sons.

Greatorex, J. (2007). *Contemporary GCSE and A-level Awarding: A psychological perspective on the decision-making process used to judge the quality of candidates' work*. Paper presented at the British Educational Research Association Conference, 5–8 September in London, UK.

Greatorex, J. & Nádas, R. (2008). *Using 'thinking aloud' to investigate judgements about A-level standards: does verbalising thoughts result in different decisions?* Paper presented at the British Educational Research Association Annual Conference, 3–6 September in Edinburgh, UK.

Kimbell, R., Wheeler, A., Miller, S. & Pollitt, A. (2007). *E-scape portfolio assessment phase 2 report*. Technology Education Research Unit, Goldsmiths College, University of London.

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

Linacre, J.M. (2005). FACETS Rasch measurement computer program. www.winsteps.com

Novaković, N. & Suto, I. (2009). *How should grade boundaries be determined in examinations? An exploration of the script features that influence expert judgements*. Paper presented at the European Conference on Educational Research, 28–30 September in Vienna, Austria.

OCR. (2008). *OCR Procedures for Awards*. Revised April 2008. Cambridge: OCR.

Raikes, N., Scorey, S. & Shiell, H. (2008). *Grading examinations using expert judgements from a diverse pool of judges*. Paper presented at the 34th Annual Conference of the International Association for Educational Assessment, 7–12 September in Cambridge, UK.

Thurstone, L.L. (1927a). Psychophysical analysis. *American Journal of Psychology*, **38**, 368–389. In: L.L. Thurstone (1959), *The measurement of values*. Chicago: University of Chicago Press.

Thurstone, L.L. (1927b). A law of comparative judgment. *Psychological Review*, **3**, 273–286.

Willmott, A.S. & Nuttall, D.L. (1975). *The reliability of examinations at 16+*. Schools Council Research Studies, Schools Council Publications. London: MacMillan Education Ltd.

How do examiners make judgements about standards? Some insights from a qualitative analysis

Jackie Greatorex Research Division

An earlier version of this article was presented at the American Educational Research Association conference, April 2009

Abstract

There is a good deal of research about how judgements are made in awarding when A level and GCSE grade boundaries are chosen. There is less research about how judgements are made in Thurstone paired comparisons and rank ordering (popular methods in comparability studies to compare grading standards). Therefore, the research question for the present study is 'how do Principal Examiners[1] (PEs) make judgements about standards in awarding, Thurstone paired comparisons and rank ordering?' The present article draws from a wider project in which Principal Examiners thought aloud whilst making judgements about the quality of candidates' work and grading standards in awarding, Thurstone paired comparisons and rank ordering situations analogous to how these methods are practised. For the present analysis a coding frame was developed to qualitatively analyse the think aloud data. The coding frame constituted codes grounded in the think aloud data and grade descriptors[2] from the qualification specification. It was found that overall the Principal Examiners attended to valid factors such as where marks were gained, responses to key questions and characteristics of candidates' work that were in the grade descriptors. When the importance of each factor was considered there were some similarities and some differences between the methods. Implications and recommendations are discussed.

Background

The focus of this article is the often asked question 'how do Principal Examiners make judgements about standards?' This question can be addressed from various perspectives including:

- What cognitive strategies do PEs use?
- What features do PEs attend to (and are they valid features)?
- What procedures are used to make decisions?

In the current article three approaches to judging grading standards are considered: (i) awarding – part of the conventional approach to recommending grade boundaries, (ii) Thurstone pairs and (iii) rank ordering. The latter two were suggested as possible future methods of

1 Principal Examiners generally write an examination question paper, lead the associated marking and take part in awarding. Most participants in Thurstone paired comparison and rank ordering studies are Principal Examiners.

2 Grade descriptors (descriptions) are written descriptions that indicate the level of attainment characteristic of a particular qualification. They give a general indication of the learning outcomes at a given grade. The descriptions should be interpreted in relation to the content outlined in specifications, they do not outline the specification content (OCR, 2004). A specification is a description of what can be tested in an examination. Note that this research was undertaken before specifications began providing **performance descriptions** rather than **grade descriptions**. Performance descriptors (descriptions) are written descriptions of the typical knowledge, skills and understanding likely to be found in candidates' work at the judgementally awarded grade boundaries. These descriptors are indicators of the knowledge, understanding and skills that are likely to be found in candidates' work at the grade boundary, they are not requirements. There might be other knowledge, understanding and skills that are found in candidates' work at the grade boundary. They are designed to aid recommending grade boundaries.

recommending grade boundaries by Pollitt and Elliott (2003a and b), Black and Bramley (2008) and Kimbell *et al*. (2007). They have also been used in a series of comparability studies (e.g. Forster and Gray, 2000; Arlett, 2003; Greatorex *et al*., 2002, 2003; Edwards and Adams, 2002, 2003; Guthrie, 2003; Bramley *et al*., 1998; Townley, 2007). Note that Thurstone pairs and rank ordering are not currently used in operational awarding or in operational procedures to recommend grade boundaries.

What are the current practices for awarding, Thurstone paired comparisons and rank ordering?

In this research the focus is on one decision-making phase of awarding which involves the awarding committee judging whether a small number of examples of candidates' work[3] on particular marks show the distinguishing characteristics of performance at a particular grade. For a fuller description, see Cresswell (1997), QCA (2008) or Greatorex (2003a).

Thurstone pairs and rank ordering have been frequently described in the literature and there are many examples of their use in comparability studies; see for example Bramley *et al*. (1998), Arlett (2003), Greatorex *et al*. (2002, 2003), Edwards and Adams (2002, 2003), Guthrie (2003) and Townley (2007). Both Thurstone pairs and rank ordering involve a group of experts judging the quality of candidates' work.

In a Thurstone pairs design each expert compares a pair of scripts. In a study investigating standards maintenance, each pair would consist of a script from the most recent examination and one from the archive examination. Each expert decides which of two scripts contains the better performance, without re-marking the scripts. This is repeated for a variety of pairs of scripts. Once all the necessary comparisons are complete, they are statistically analysed (using Rasch). The results of the analysis can be used to identify a small range of marks within which the most recent boundary should lie for the standard from last year to be maintained.

In a study investigating standards maintenance using a rank ordering design each expert is given small samples of scripts which they rank from best to worst performance. Each small sample has a mixture of most recent and archive scripts. This is repeated for a number of overlapping samples of scripts. The outcomes of the rankings are submitted to the same statistical analysis as above. Again the statistics can be used to identify a small range of marks within which the most recent boundary should lie.

What does research tell us about how judgements are made about grading standards?

There is a good deal of research about judgements of grading standards in awarding, for example, Good and Cresswell (1988a and b), Scharaschkin and Baird (2000), Baird and Scharaschkin (2002). The present literature review will be confined to the most relevant literature.

Murphy *et al*. (1995) argue that each awarding committee member's impressions of what was appropriate were from a variety of sources, three of which were identified in their research:

1. knowledge of requirements of the national curriculum or other descriptions of performance;
2. performance on questions that some believed to be indicative of achievement (and the belief that it was possible to make judgements on these alone);
3. the belief that they 'knew' what constitutes work at a particular grade.

They found that the general use of archive material was low. Later Baird (2000) found that the severity of judgements of grade-worthiness was sometimes influenced by the archives provided. Research shows archive scripts were sometimes missed in awarding in the past. Archive scripts are still a useful source of information listed in the Code of Practice.

Cresswell (1997) investigated the weighting of many factors in judgements about grading standards such as technical and statistical evidence as well as the features noted in candidates' work. He found little evidence that the demand of questions was taken into account when PEs judged the candidates' work. Cresswell (1997) and later Crisp (2007, 2008) found that valid features of candidates' work contributed to decisions about grading standards. Crisp (2008) found that PEs made judgements by paying attention to features in candidates' work which were closely tied to the mark scheme, such as a good understanding of concepts, application of knowledge and evaluation and application of skills. However, Cresswell (1997) also argued that other less valid features also had some input in judgements of grading standards. For example, sometimes features such as whether the candidate's work gave the reader pleasure or was interesting were taken into account, when they were not necessarily linked to the features intended to be judged (Cresswell, 1997).

There are various aspects of awarding meetings and scripts that positively and negatively influence judgements of gradeworthiness (Cresswell, 1997; Murphy *et al*., 1995; Crisp, 2007; Baird, 2000; Baird and Scharaschkin, 2002; Scharaschkin and Baird, 2000). To consider this further it is important to note that A level and GCSE examinations have a principle of compensation, according to which candidates gain marks for their strengths, and there is more than one way to achieve a grade. Two conundrums relate to the principle of compensation and the visibility of marks on scripts:

- Some PEs in some awarding meetings particularly focus on questions and marks which are believed to differentiate between performances at particular grades (Murphy *et al*., 1995; Greatorex *et al*., 2008). This belief might be well or ill founded (Murphy *et al*., 1995). Focussing on particular questions at the expense of other questions is not aligned with the principle of compensation. Psychological research from a variety of contexts suggests that humans are not particularly good at combining information to make decisions. For a detailed discussion of this, see Greatorex (2007) and Greatorex *et al*. (2008). Therefore, focussing judgements on particular questions might be a successful approach to decision making, if the questions are a good proxy for the whole of the examination. After all, the other strategy – judgements about whole scripts – involves mentally combining a candidate's answers to all questions in the examination.
- It has been established that the consistency of candidates' performance across questions on an examination paper influences the severity of judgements of gradeworthiness (Cresswell, 1997; Scharaschkin and Baird, 2000). Again, this is not aligned with the principle of compensation.

3 The candidates' work is usually written examination scripts but might be a recording of a drama or musical performance or an artefact such as a painting.

There is a small amount of research about how judgements are made in Thurstone paired comparisons comparability studies. For example, Edwards and Adams (2002, 2003) asked PEs in Thurstone paired comparison studies what criteria they used to judge the candidates' work. They report that the criteria were quite wide ranging, but that some of the common criteria included "depth of understanding" and "level of reasoning" (Edwards and Adams 2003, p.20). All the examples that they list seem to be valid and reasonable criteria for judging the candidates' work. This reassures us that for some Thurstone paired comparison studies judgements are made by taking valid information into account. In rank ordering studies the correlation between the trait 'perceived quality of candidates' work' and total mark is pleasingly high (between 0.8 and 0.9) (Bramley, 2007). Thus we have some evidence that rank ordering is measuring something similar to the total marks, and that the judgements are valid.

Context of the present study

The present study is the third in a series of inter-linked studies which draw from a wider research project. The research is still in progress. The first and second studies are reported in Greatorex and Nádas (2009) and Greatorex *et al*. (2008). In the wider project the aim is to find out more about cognitive processes used by PEs to make judgements about grading standards.

Greatorex and Nádas (2009) found that, broadly speaking, the task outcomes were similar whether the judgements were made silently or whilst thinking aloud. Therefore, there is some evidence that research results using the think aloud data are trustworthy.

Greatorex *et al*. (2008) studied which examination question responses or answers the PEs referred to in the candidates' work. They found that the questions most often referred to did not always discriminate well between achievements just above and below the grade boundary. This ties in with Murphy *et al*.'s (1995) concern that the questions used as key discriminators might or might not statistically discriminate between performances on the two adjacent grades. Therefore, the Research Division at Cambridge Assessment argued that question level data from on-screen marking should be used to facilitate choosing key discriminating questions.

Thus far the reporting of the wider research project, of which this study forms a part, has covered a quantitative analysis of the outcomes of the tasks, and qualitative coding using *a priori* codes (the examination questions). What has not been reported is a qualitative analysis using codes that are grounded in the rich textual content of the think aloud data, and therefore this is the focus of the present study.

Method

The method for the project is reported in more detail in Greatorex and Nádas (2009). Two past AS biology examinations were used as a source of data. The first year of the examination will be referred to as the 'archive examination' and the next year of the examination will be referred to as the 'live examination'. The five participants (called PEs in this report) had all been involved in awarding the AS examination. All the examples of candidates' work used in the research were from near the grade boundaries from the two examinations.

Prior to the main data collection phase PEs undertook some warm up exercises including:

- Thinking aloud whilst doing non-examining tasks.
- Silently making decisions in the five experimental conditions described below.

In the main data collection phase PEs thought aloud whilst making judgements in the five experimental conditions:

- Awarding with marks visible ('awarding visible');
- Awarding with candidates' work cleaned of marks ('awarding clean');
- Thurstone paired comparisons with marks visible ('Thurstone visible');
- Thurstone paired comparisons with candidates' work cleaned of marks ('Thurstone clean');
- 'Rank ordering' with candidates' work cleaned of marks.

The thinking aloud was audio recorded and transcriptions were made.

The awarding conditions reflected the aspect of awarding where individual committee members evaluate scripts, before coming to a collective view about where the grade boundary should be. The rank ordering and Thurstone pairs conditions were intended to reflect current/best practices in prior studies. For all experimental conditions some small adjustments were made to current/best practices for the purposes of this research. Photocopies of the scripts were used rather than the original scripts. For each method the scripts were presented as they are normally presented: awarding with marks visible and rank ordering with scripts cleaned of marks. Thurstone pairs studies vary regarding whether the marks are visible or not so this was reflected in the research. 'Awarding clean' reflected the aspect of awarding where individual awarding committee members evaluate scripts, before coming to a collective view about where the grade boundary should be. But in 'awarding clean' the scripts were cleaned of marks. A reason for this experimental control was the arguably extraneous influence of visible marks in some awarding judgements (Murphy *et al*., 1995; Cresswell, 1997; Scharaschkin and Baird, 2000).

The script samples for the decisions made whilst thinking aloud constituted scripts with total marks within the range of marks considered in the recommendation for the grade A boundary in the awarding meeting (33 to 37 for 2005 and 28 to 34 for 2006). The live grade A boundary was 35 marks for the 2005 examination and 31 marks for the 2006 examination.

Coding for the present study

The present study involved developing a coding frame to qualitatively analyse the think aloud data. The coding frame constituted codes grounded in the think aloud data and grade descriptors from the qualification specification. To develop the coding frame the transcripts, instructions to PEs, examples of candidates' work and grade descriptors were read. Although the grade descriptors are not used in the grading process, it is likely that they would give a good indication of senior examiners' views of achievement at each grade. Over a series of iterations of reading and trying out codes and coding frames, a coding frame grounded in the data was developed. The process was informed by some of the content of the transcripts as well as anecdotal conversations with the PEs.

The final coding frame is described in Table 1 and Table 2. Some codes were used to identify when examiners paid attention to responses to

Table 1: The coding frame of codes grounded in the think aloud data and the question papers

Shorthand label used in coding the transcripts	*What the question(s) required candidates to do/topics tested*	*What the PEs seem to be doing*
'Archive/ Question A'	Explain the significance of the different affinities of foetal haemoglobin and adult haemoglobin for oxygen	The PEs seemed to consider this question to be a high demand question and therefore a good source of information about A and B grade performance.
'Archive/ Question B'	Explain the significance of the dissociation curve of adult haemoglobin	As above
'Comparing long answers'	Explain the relationship between the structure and function of arteries, veins and capillaries	One question in each examination was a long answer question on a somewhat similar topic so sometimes the answers from different years were compared or referred to.
'Live/Question X'	Explain translocation as an energy requiring process	The PEs seemed to consider this question to be a high demand question and therefore a good source of information about A and B grade performance.
'Live/Question X–'	Explain translocation as an energy requiring process	The PEs seemed to consider this question to be a high demand question and therefore a good source of information about A and B grade performance. This code applied only to negative comments about the candidates' work.
'Live/Question X+'	Explain translocation as an energy requiring process	The PEs seemed to consider this question to be a high demand question and therefore a good source of information about A and B grade performance. This code applied only to positive comments about the candidates' work.
'Live/Question Y'	Describing the mammalian circulatory system as a closed double circulation	Question Y in the live examination was arguably a lower demand question than those listed above but seems to have been seen as a good source of information.

Table 2: The coding framework of codes grounded in the think aloud data and the mark scheme or grade descriptors

Shorthand label used in coding the transcripts	*What the PE seems to be doing*
'Explain'	The PE seems to be looking for a characteristic listed in the grade descriptor, i.e. provide coherent and logical explanations.
'Identify marks'	The PE seems to be trying to identify where marks were given.
'Know and understand'	The PE seems to be looking for a characteristic listed in the grade descriptor, i.e. show good knowledge and understanding.
'Present'	The PE seems to be looking for a characteristic listed in the grade descriptor, i.e. present ideas clearly and logically.

particular items, these are given in Table 1. Other codes were grounded in the protocols, mark scheme and grade descriptors (see Table 2). Each code was taken to be a factor that contributed to judgements about grading standards.

Unfortunately, for some PEs there was not time to complete all the tasks and in places transcripts are ambiguous, resulting in some missing data.

A sample of data was double-coded. The second coder did not see the original coding. Once the double-coding was collated, only the most reliably coded codes were retained.

Once the coding was complete, it was established which code(s) was present in the section of the transcript associated with each example of candidate's work. Next, the presence data was expressed as a proportion of the total number of examples of candidates' work available in each condition for all PEs. For instance, the following is a hypothetical example: there were 100 examples of candidates' work in total in 'rank ordering'. Code A was present for candidates 1 to 60, so code A had a proportion of 60%, whereas, code B was present only for candidates 5 and 6, and so had a proportion of 2%. The proportions were ranked in descending order. Therefore, the higher the rank, the more important the code (or associated factor) is in making judgements. Using our example the factor associated with code A was more important in making judgements than the factor associated with code B. A limitation of this analysis is that some information is lost by ranking rather than using frequencies or similar.

Results

Overall, the PEs made judgements in all the conditions by paying attention to:

- Responses about particular areas of content (questions) which seemed to be perceived as a good source of information about A and B grade performance and/or were perceived to be high demand questions.
- Responses to the long answer question in each examination which had some overlap in the subject content tested, and therefore seemed a solid basis for comparison between the performance in the two different examinations.
- Some characteristics referenced in the grade descriptors.
- Whether the candidates seemed to have been credited with marks.

This is summarised in Figure 1.

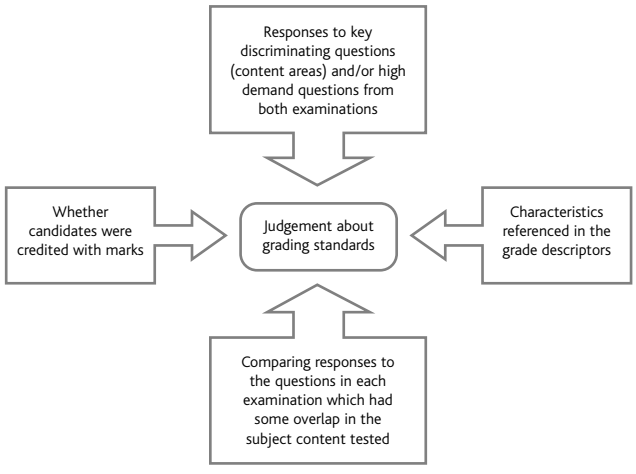


Figure 1: The overarching themes that contributed to judgements

In addition to the overarching themes that contributed to judgements about grading standards there were the factors identified in the coding frame. The following text boxes give the rank of the importance of each factor in judgements for each condition. Note that some of the ranks are ties and therefore some ranks are repeated and others are omitted. For example, for 'awarding visible' two codes were ranked 9 and no codes were ranked 10.

How were judgements made in 'awarding visible'?

Overall judgements were made by looking for correct answers, focussing on answers to particular questions or areas of the syllabus and looking for characteristics listed in grade descriptors. In descending order of importance the factors taken into account were:

1 Responses to question Y in the live examination. This question was about the mammalian circulatory system and seemed to be considered a good source of information about A/B grade performance.

2 Identifying where marks were given.

3 Comparing responses to the long answer questions (one in each examination) which were both about explaining the relationship between the structure and function of arteries, veins and capillaries. These questions seemed to be perceived as a good source of information about grade A/B performance.

4 Negative views about performance on question X in the live examination. This question expected candidates to explain translocation as an energy requiring process.

5 Positive views about performance on question X in the live examination.

6 Responses to question B in the archive examination. This question expected candidates to explain the significance of the dissociation curve of adult haemoglobin and seemed to be viewed as a good source of information about grade A/B performance.

7 Finding evidence of a characteristic in the grade descriptors; in this case 'presents ideas clearly and logically'.

8 Neutral views about performance on question X in the live examination.

9 Finding evidence of characteristics in the grade descriptors; in this case 'provides coherent and logical explanations'.

9 Finding evidence of characteristics in the grade descriptors; in this case 'shows good knowledge and understanding'.

11 Responses to question A in the archive examination. This question expected candidates to explain the significance of the different affinities of foetal haemoglobin and adult haemoglobin for oxygen and seemed to be viewed as a good source of information about grade A/B performance.

How were judgements made in 'awarding clean'?

Overall judgements were made by looking for correct answers, focussing on answers to particular questions and looking for characteristics listed in grade descriptors. In descending order of importance the factors taken into account were:

1 Responses to question Y in the live examination. This question was about the mammalian circulatory system and seemed to be considered a good source of information about A/B grade performance.

2 Identifying where marks were given.

3 Comparing responses to the long answer questions (one in each examination) which were both about explaining the relationship between the structure and function of arteries, veins and capillaries. These questions seemed to be perceived as a good source of information about grade A/B performance.

4 Positive views about performance on question X in the live examination. This question expected candidates to explain translocation as an energy requiring process.

5 Responses to question A in the archive examination. This question expected candidates to explain the significance of the different affinities of foetal haemoglobin and adult haemoglobin for oxygen and seemed to be viewed as a good source of information about grade A/B performance.

6 Negative views about performance on question X in the live examination.

7 Finding evidence of a characteristic in the grade descriptors; in this case 'provides coherent and logical explanations'.

8 Responses to question B in the archive examination. This question expected candidates to explain the significance of the dissociation curve of adult haemoglobin and seemed to be viewed as a good source of information about grade A/B performance.

9 Neutral views about performance on question X in the live examination.

10 Finding evidence of a characteristic in the grade descriptors; in this case 'presents ideas clearly and logically'.

11 Finding evidence of characteristics in the grade descriptors; in this case 'shows good knowledge and understanding'.

How were judgements made in 'Thurstone clean'?

Overall judgements were made by looking for correct answers, focussing on answers to particular questions and looking for characteristics listed in grade descriptors. In descending order of importance the factors taken into account were:

1 Identifying where marks were given.

2 Comparing responses to the long answer questions (one in each examination) which were both about explaining the relationship between the structure and function of arteries, veins and capillaries. These questions seemed to be perceived as a good source of information about grade A/B performance.

2 Responses to question Y in the live examination. This question was about the mammalian circulatory system and seemed to be considered a good source of information about A/B grade performance.

4 Responses to question A in the archive examination. This question expected candidates to explain the significance of the different affinities of foetal haemoglobin and adult haemoglobin for oxygen and seemed to be viewed as a good source of information about grade A/B performance.

5 Responses to question B in the archive examination. This question expected candidates to explain the significance of the dissociation curve of adult haemoglobin and seemed to be viewed as a good source of information about grade A/B performance.

6 Finding evidence of a characteristic in the grade descriptors; in this case 'provides coherent and logical explanations'.

7 Positive views about performance on question X in the live examination. This question expected candidates to explain translocation as an energy requiring process.

7 Negative views about performance on question X in the live examination.

9 Neutral views about performance on question X in the live examination.

10 Finding evidence of a characteristic in the grade descriptors; in this case 'shows good knowledge and understanding'.

11 Finding evidence of a characteristic in the grade descriptors; in this case 'presents ideas clearly and logically'.

How were judgements made in 'Thurstone visible'?

Overall judgements were made by looking for correct answers, focussing on answers to particular questions and looking for characteristics listed in grade descriptors. In descending order of importance the factors taken into account were:

1 Identifying where marks were given.

2 Comparing responses to the long answer questions (one in each examination) which were both about explaining the relationship between the structure and function of arteries, veins and capillaries. These questions seemed to be perceived as a good source of information about grade A/B performance.

2 Responses to question A in the archive examination. This question expected candidates to explain the significance of the different affinities of foetal haemoglobin and adult haemoglobin for oxygen and seemed to be viewed as a good source of information about grade A/B performance.

4 Responses to question B in the archive examination. This question expected candidates to explain the significance of the dissociation curve of adult haemoglobin and seemed to be viewed as a good source of information about grade A/B performance.

5 Responses to question Y in the live examination. This question was about the mammalian circulatory system and seemed to be considered a good source of information about A/B grade performance.

6 Finding evidence of a characteristic in the grade descriptors; in this case 'presents ideas clearly and logically'.

7 Negative views about performance on question X in the live examination. This question expected candidates to explain translocation as an energy requiring process.

7 Positive views about performance on question X in the live examination.

9 Finding evidence of a characteristic in the grade descriptors; in this case 'shows good knowledge and understanding'.

10 Neutral views about performance on question X in the live examination.

11 Finding evidence of a characteristic in the grade descriptors; in this case 'provides coherent and logical explanations'.

How were 'rank ordering' judgements made?

Overall judgements were made by looking for correct answers, focussing on answers to particular questions and looking for characteristics listed in grade descriptors. In descending order of importance the factors taken into account were:

1 Identifying where marks were given.

2 Responses to question A in the archive examination. This question expected candidates to explain the significance of the different affinities of foetal haemoglobin and adult haemoglobin for oxygen and seemed to be viewed as a good source of information about grade A/B performance.

3 Comparing responses to the long answer questions (one in each examination) which were both about explaining the relationship between the structure and function of arteries, veins and capillaries. These questions seemed to be perceived as a good source of information about grade A/B performance.

4 Responses to question B in the archive examination. This question expected candidates to explain the significance of the dissociation curve of adult haemoglobin and seemed to be viewed as a good source of information about grade A/B performance.

5 Responses to question Y in the live examination. This question was about the mammalian circulatory system and seemed to be considered a good source of information about A/B grade performance.

6 Negative views about performance on question X in the live examination. This question expected candidates to explain translocation as an energy requiring process.

7 Finding evidence of a characteristic in the grade descriptors; in this case 'provides coherent and logical explanations'.

7 Positive views about performance on question X in the live examination.

9 Neutral views about performance on question X in the live examination.

10 Finding evidence of a characteristic in the grade descriptors; in this case 'presents ideas clearly and logically'.

11 Finding evidence of characteristics in the grade descriptors; in this case 'shows good knowledge and understanding'.

There are some commonalities in the importance of the different factors in the judgements made in different conditions (see Table 3). 'Identify marks' was ranked amongst the two most important factors for all conditions, and 'comparing long answers' was in the top three most important factors for all conditions. Additionally, 'know and understand' (showing good knowledge and understanding) was ranked amongst the three least important factors for all conditions. 'Live/question X' was also ranked amongst the four least important factors for all conditions.
There were also some differences in the rank order of importance of the different factors in different conditions (see Table 3). Factor 'archive/question A' was ranked in the top five most important factors for the 'awarding clean', 'rank ordering', 'Thurstone visible' and 'Thurstone clean' conditions, but was ranked as the least important factor for 'awarding visible'. Factor 'live/question Y' was ranked in the top two most important factors for the 'awarding clean', 'awarding visible' and 'Thurstone clean' conditions, but was ranked as lower for the 'rank ordering' and 'Thurstone visible' conditions. Factor 'live/question X+' was ranked as fourth most important for the 'awarding clean' condition but as low as seventh for the 'rank ordering', 'Thurstone clean' and the 'Thurstone visible' conditions. The factor 'present' was ranked as low (tenth or lower) for the 'Thurstone clean', 'awarding clean' and 'rank ordering' conditions, but seventh or higher for the 'awarding visible' and 'Thurstone visible' conditions.

Discussion

The main research question for the present study is 'how do Principal Examiners make judgements about grading standards in awarding, Thurstone paired comparisons and rank ordering?' It was found that overall the PEs attended to valid factors such as where marks were gained, responses to key questions and characteristics of candidates' work that are referenced in the grade descriptors. This finding was apparent for all conditions, and might be somewhat generalisable to the methods – awarding, Thurstone paired comparisons and rank ordering. When the importance of each factor was considered there were some similarities and some differences between the methods.

There are a number of limitations to the present study. First, it is not possible to generalise about all GCSE and A-level judgements of grading standards from two examinations and one judgementally awarded grade

Table 3: The rank order of importance of each factor in judgements of grading standards

Shorthand label	*rank order of importance in judgements*				
	'awarding visible'	*'awarding ' clean*	*'rank ordering'*	*'Thurstone visible'*	*'Thurstone clean'*
'Archive/question A'	11	5	2	2 =	4
'Archive/question B'	6	8	4	4	5
'Comparing long answers'	3	3	3	2 =	2 =
'Live/question X'	8	9	9	10	9
'Live/question X–'	4	6	6	7 =	7 =
'Live/question X+'	5	4	7 =	7 =	7 =
'Live/question Y'	1	1	5	5	2 =
'Explain'	9 =	7	7 =	11	6
'Identify marks'	2	2	1	1	1
'Know and understand'	9 =	11	11	9	10
'Present'	7	10	10	6	11

Note: 1 is the highest rank; = denotes ties

boundary. However, the examinations were carefully chosen as examinations which might involve judgements about numerical skills, written skills, use of diagrams, and knowledge and understanding, whereas in some other subjects PEs might judge candidates' work which is predominantly in one skill area. Secondly, only think aloud was used as a method of data collection. It is often advised that think aloud data are used to generate hypotheses which are tested out in further empirical studies. To this end there is research underway at Cambridge Assessment to identify which features of candidates' work are used in judgements about grading standards using a more quantitative and generalisable approach. Thirdly, the 'awarding clean' and 'awarding visible' conditions have limited ecological validity; they do not include much of the information that is available in traditional awarding meetings, and they omit the face to face social dynamics of the awarding meeting. For research that incorporates these influences see Murphy *et al*. (1995) and Cresswell (1997). However, the awarding meeting information was not provided to avoid it influencing the judgements in the other conditions. Furthermore, if remote awarding becomes more widespread then there might be an increase in individual decision making which reflects the think aloud setting in this study when a PE made judgements without other PEs present.

The general themes that the PEs attended to (characteristics referenced in the grade descriptors, key discriminating questions, comparing answers to similar questions from different years of the examination and identifying where marks were given) all seem to be valid sources of information for making judgements about grading standards. The limitations and strengths of using key discriminating questions have been considered by Murphy *et al*. (1995) and later by Greatorex *et al*. (2008). For example, more credit might be given to responses to particular questions than was intended by the mark scheme. Additionally, it is important that the question is measuring the same as the whole examination. Comparing answers to similar questions from the two years of the examination shares the strengths and limitations of using key discriminating questions. The finding that PEs attend to some specific items, and that the items seem to be used because of the demands they place on candidates, illustrates that the context in which the candidates perform is important to PEs' decision making. This is a contrast to Cresswell's finding that the PEs did not pay much attention to the demands of the questions and how this affected candidates' performance.

Much of the previous literature has suggested that PEs compare the candidates' work with their impression of what is appropriate to a particular grade (sometimes referred to as a prototype or internal standard) (Murphy, 1995; Baird, 2000; Crisp, 2008). In the present analysis it was found that PEs attended to features referenced in the grade descriptors. In line with current awarding practices the grade descriptors were not available during the thinking aloud and therefore the PEs must have been remembering them, or the descriptors are a good reflection of the prototypes that PEs have for performance at grades A and B. This ties in with the well-rehearsed argument that grade descriptors should be grounded in both candidates' actual performance and Principal Examiners' views of the features that discriminate between achievement at different grades (Greatorex, 2001, 2002, 2003b; Greatorex *et al*., 2001). PEs seem to be looking for particular features and using particular features in judgements whether they are comparing the candidates' performance with a prototype, or with a memory of another candidate's work.

Crisp (2007) and Bramley (2007) indicate that there is commonality between what is given credit in the mark scheme (measured by total mark) and what contributes to judgements of grading standards. This ties in with the finding in the present analysis that PEs try to identify what marks were given.

The general themes which contributed to judgements of grading standards reflect some of the existing literature. However, what has not previously been reported is a comparison of the judgement process in awarding versus Thurstone paired comparisons versus rank ordering, and this is the focus of the next section.

There were some commonalities between the factors that were ranked as the most and least important factors in making judgements. For instance, 'comparing long answers' was ranked high for all conditions, and this corroborates the findings of Greatorex *et al*. (2008). Therefore, it seems that there are some commonalities in how PEs make judgements in each of the conditions. On the other hand there were also some differences in the rank of importance of the different factors in different conditions. There was no clear overall pattern regarding whether two or more conditions were particularly similar in how PEs made judgements.

There were some differences in the rank order of importance of the various factors in different conditions. The factor 'present' was ranked as low (tenth or lower) for the 'Thurstone clean', 'awarding clean' and 'rank ordering' conditions, but seventh or higher for the 'awarding visible' and 'Thurstone visible' conditions. Also 'archive/question A' and 'archive question B' were ranked lower in 'awarding visible' and in 'awarding clean', than in the comparability study conditions. This appears to somewhat corroborate Murphy *et al*'s (1995) finding that the archive scripts are infrequently used, however, awarding practices have changed since their work and the Code of Practice (2008, p36) says that the archive "must be used, as appropriate, to inform the determination of marks at key grade boundaries". Indeed Laming's (2004) work about humans being better at making comparisons than maintaining internal standards would suggest that as far as possible awarding procedures should recommend systematic

and frequent comparisons between the archive and the live examples of candidates' work. It is not clear why the importance of some other factors varies between conditions. For example, 'live/question Y' is amongst the two most important factors in the 'awarding visible', 'awarding clean' and 'Thurstone clean' but is of lower ranking in the other conditions.

Previous research has tended to compare the trait measured in comparability studies with total marks rather than awarding judgements; see for example Bramley (2007). However, the present study offers the opportunity to compare what might be measured in comparability studies with what is measured in awarding. This is accomplished by treating what PEs attend to as a strong proxy for what is measured. The present study suggests the trait 'perceived quality of candidates' work' might vary a little with the condition that is used in comparability studies (rank ordering or Thurstone paired comparisons), and might also differ somewhat from what is measured in awarding at a particular boundary. However, as explained earlier there are also strong commonalities between conditions regarding both the factors PEs attend to and their importance in judgements. If there were system changes as suggested by Pollitt and Elliott (2003a and b) or Black and Bramley (2008) then what is being measured might change slightly. However, in all approaches in this research PEs attended to valid factors, so what was measured when using each method is arguably valid.

The present study has offered many insights into what PEs attend to when they make the judgements about grading standards, from psychological and other perspectives. However, it is somewhat difficult to generalise from this particular analysis to other examinations, as some of the coding refers to aspects of biology. The next stage in the wider research project is to undertake a more psychological analysis with particular focus on whether PEs are making comparisons between candidates' work or whether they are using internal standards.

Acknowledgements

Thank you to the senior examiners who took part in this research. Thank you also to Richard Shewry who provided confidential consultancy for this research project; amongst other things he offered insights from the biology teaching and assessment community which is not the specialist area of the main researcher from this project.

References

Arlett, S.J. (2003). *A comparability study in VCE Health and Social Care, Units 3, 4 and 6: a review of the examination requirements and a report on the cross moderation exercise*. A study based on the Summer 2002 examination and organised by AQA on behalf of the Joint Council for General Qualifications.

Baird, J. (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. & 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, 143–162.

Black, B., & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education*, **23**, 3, 357–373.

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgement. *Journal of Applied Measurement*, **6**, 2, 202–223.

Bramley, T. (2007). Paired Comparison Methods. 246–294 In: P Newton, J Baird, H Goldstein, H Patrick and P Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. QCA: London.

Bramley, T., Bell, J.F. & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone paired comparisons. *Education Research and Perspectives*, **25**, 2, 1–23.

Cresswell, M. (1997). *Examining Judgements: Theory and Practice of Awarding public examination grades*. PhD thesis, University of London Institute of Education: London.

Crisp, V. (2007). *Do assessors pay attention to appropriate features of student work when making assessment judgements?* A paper presented at the International Association for Educational Assessment Conference, Baku, Azerbaijan, September 2007.

Crisp, V. (2008). *Judging the grade: An exploration of the judgement processes involved in A-level grading decisions*. A paper presented at the British Educational Research Association Conference, Heriot Watt University, Edinburgh, September 2008.

Edwards, E. & Adams, R. (2002). *A Comparability Study in GCE Advanced Level Geography including the Scottish Advanced Higher Grade Examinations. A review of the examination requirements and a report on the cross moderation exercise*. A study based on the Summer 2001 Examination and organised by WJEC on behalf of the Joint Council for General Qualifications.

Edwards, E. & Adams, R. (2003). *A Comparability Study in GCE Advanced Level Geography including the Scottish Advanced Higher Grade Examinations. A review of the examination requirements and a report on the cross moderation exercise*. A study based on the Summer 2002 Examination and organised by WJEC on behalf of the Joint Council for General Qualifications.

Forster, M. & Gray, E. (2000) *Impact of Independent Judges in comparability studies conducted by Awarding Bodies*. A paper presented at the British Educational Research Association Annual Conference, Cardiff University, Cardiff, September 2000.

Good, F. J. & Cresswell, M. J. (1988a). Grade Awarding Judgements in differentiated examinations. *British Educational Research Journal*, **14**, 3, 263–281.

Good, F. J. & Cresswell, M. J. (1988b). Grading the GCSE. London: Secondary Schools Examination Council. In: M Cresswell (2000), Defining, Setting and Maintaining Standards in Curriculum-Embedded Examinations: Judgemental and Statistical Approaches. In: H Goldstein, and T Lewis (Eds.) *Assessment: Problems, developments and statistical issues*. 57–84. John Wiley and Sons: Chichester.

Greatorex, J. (2001). Making the grade – how question choice and type affect the development of grade descriptors. *Educational Studies*, **27**, 4, 451–464.

Greatorex, J. (2002). Making Accounting examiners' tacit knowledge more explicit: developing grade descriptors for Accounting A-Level. *Research Papers in Education*, **17**, 2, 211–226.

Greatorex, J. (2003a). *What happened to limen referencing? An exploration of how the Awarding of public examinations has been and might be conceptualised*. A paper presented at the British Educational Research Association Conference, Heriot Watt University, Edinburgh, September 2008.

Greatorex, J. (2003b). Developing and applying level descriptors. *Westminster Studies in Education*, **26**, 2, 125–133.

Greatorex, J. (2007). *Contemporary GCSE and A-level Awarding: A psychological perspective on the decision-making process used to judge the quality of candidates' work*. A paper presented at the British Educational Research Association Conference, University of London, London, September 2007.

Greatorex, J. & Nádas, R. (2009). Using 'thinking aloud' to investigate judgements about A-level standards: does verbalising thoughts result in different decisions? *Research Matters: A Cambridge Assessment Publication*, **7**, 8–16. Also presented at the British Educational Research Association Conference, Heriot Watt University, Edinburgh, September 2008.

Greatorex, J., Elliott, G. & Bell, J. F. (2002). *A comparability study in GCE AS Chemistry including parts of the Scottish Higher Grade Examinations, A review of the examination requirements and a report on the cross moderation exercise*. A study based on the Summer 2001 examination and organised by the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for OCR on behalf of the Joint Council for General Qualifications.

Greatorex, J., Hamnett, L. & Bell, J.F. (2003). *A comparability study in GCE Chemistry including the Scottish Advanced Higher Grade*. A study based on the Summer 2002 examination and organised by the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for OCR on behalf of the Joint Council for General Qualifications.

Greatorex, J., Johnson, C. & Frame, K. (2001). Making the grade – developing grade profiles for accounting using a discriminator model of performance. *Westminster Studies in Education*, **24**, 2, 167–181.

Greatorex, J., Novaković, N. & Suto, I. (2008). *What attracts judges' attention? A comparison of three grading methods*. A paper presented at the International Association for Educational Assessment Conference, Cambridge, September 2008.

Guthrie, K. (2003). *A Comparability Study in GCE Business Studies and VCE Business, A review of the examination requirements and a report on the cross moderation exercise*. A study based on the Summer 2002 Examination and organised by the Edexcel on behalf of the Joint Council for General Qualifications.

Kimbell, R., Wheeler, A., Miller, S. & Pollitt, A. (2007). *E-scape portfolio assessment phase 2 report*. Department of Design, Goldsmiths, University of London. [online.] Available at: http://www.goldsmiths.ac.uk/teru/UserFiles/File/e-scape2.pdf

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

Murphy, R., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J. & Gower, R. (1995). *The Dynamics of GCSE Awarding*. Report of a project conducted for the School Curriculum and Assessment Authority. School of Education, University of Nottingham.

OCR (2004). OCR AS GCE Business Studies (3811) OCR Advanced GCE in Business Studies (7811) Approved Specification, Revised Edition. OCR. [online.] Available at: http://www.ocr.org.uk/qualifications/as_alevelgce/business_studies/documents.html

Pollitt, A. & Elliott, G. (2003a). Monitoring and investigating comparability: a proper role for human judgement. Qualifications and Curriculum Authority Seminar on 'Standards and Comparability', UCLES, 4th April 2003.

Pollitt, A. & Elliott, G. (2003b). Finding a proper role for human judgement in the examination system. Qualifications and Curriculum Authority Seminar on 'Standards and Comparability', UCLES 4th April 2003.

Qualifications and Curriculum Authority (2008). GCSE, GCE, and AEA code of practice 2008. QCA: London.

Scharaschkin, A. & Baird, J. (2000). The effects of consistency of performance on A Level examiners' judgements of standards. *British Educational Research Journal*, **26**, 3, 343–357.

Townley, C. (2007). *Australian Education Systems Officials Committee – Secondary Schools Reporting – A study to examine the feasibility of a common scale for reporting all senior secondary subject results*. Victorian Curriculum and Assessment Authority: Melbourne.

'Key discriminators' and the use of item level data in awarding

Tom Bramley Research Division

Introduction

As more examination papers in general qualifications (GCSEs and A levels) are scanned and marked on screen, the marks on individual questions or question parts are collected automatically, and are referred to as item level data (ILD). The analysis of ILD is available for use in awarding meetings (where the grade boundaries are decided). This article discusses the theoretical rationale for using ILD in awarding, presents some possible formats for displaying data, and suggests ways in which the data could be used in practice.

For many examinations (whether marked on screen or not), the Principal Examiner (PE) will have produced a list of the questions which they expected to be 'key discriminators' at particular grade boundaries. This information might come from the test blueprint (for example, if each question on a test was 'targeted' at pupils at a particular grade or level), or it might come from the PE's (and their marking team's) experience of marking the papers – for example, if during the course of marking the paper they noticed which questions seemed to be discriminating well at particular grades or levels.

The (often unspoken) assumption behind identifying these 'key discriminators' is that by focussing on performance on these questions when making judgements about scripts in the awarding meeting, the awarding panel will use their time and effort most efficiently and be best able to identify the overall score on the test which represents the same performance standard as the corresponding grade boundary set in previous sessions.

The Guttman pattern – an idealised scenario

Imagine that we have a test consisting of ten dichotomous items (items scored 1 or 0). The scores on such a test fit a Guttman[1] pattern if success on an item implies success on all easier items and failure on an item implies failure on all harder items. If the columns represent the items with the easiest item at the left and the hardest item at the right, and the rows represent examinees with the least able at the top and the most able at the bottom, then a Guttman pattern for scores of 23 examinees on this 10-item test might look like Table 1 below.

If the score data fit this idealised pattern then all scripts on the same test total would show exactly the same performance (in terms of which items were answered correctly and incorrectly). In other words, every script perfectly represents the performance of all examinees with the same test score. Furthermore, there is a 'simple order' in the raw scores. Each increasing test total implies that the examinee has achieved

1 Louis Guttman (1916–1987) was an American psychologist. See http://en.wikipedia.org/wiki/Guttman_scale for more information.

Table 1: Illustration of Guttman pattern of test scores

	Easy →									*Hard*
	Q4	*Q2*	*Q1*	*Q7*	*Q5*	*Q8*	*Q3*	*Q10*	*Q9*	*Q6*	*Total*
E01	0	0	0	0	0	0	0	0	0	0	0
E02	1	0	0	0	0	0	0	0	0	0	1
E03	1	1	0	0	0	0	0	0	0	0	2
E04	1	1	0	0	0	0	0	0	0	0	2
E05	1	1	1	0	0	0	0	0	0	0	3
E06	1	1	1	0	0	0	0	0	0	0	3
E07	1	1	1	0	0	0	0	0	0	0	3
E08	1	1	1	1	0	0	0	0	0	0	4
E09	1	1	1	1	0	0	0	0	0	0	4
E10	1	1	1	1	0	0	0	0	0	0	4
E11	1	1	1	1	0	0	0	0	0	0	4
E12	1	1	1	1	1	0	0	0	0	0	5
E13	1	1	1	1	1	0	0	0	0	0	5
E14	1	1	1	1	1	1	0	0	0	0	6
E15	1	1	1	1	1	1	0	0	0	0	6
E16	1	1	1	1	1	1	0	0	0	0	6
E17	1	1	1	1	1	1	1	0	0	0	7
E18	1	1	1	1	1	1	1	0	0	0	7
E19	1	1	1	1	1	1	1	0	0	0	7
E20	1	1	1	1	1	1	1	1	0	0	8
E21	1	1	1	1	1	1	1	1	0	0	8
E22	1	1	1	1	1	1	1	1	1	0	9
E23	1	1	1	1	1	1	1	1	1	1	10

everything that examinees with a lower test total have achieved, plus one more item correct.

In a situation like this, the task of the award meeting would be to decide on the pattern of performance which was worthy of the particular grade – and this could be done by considering individual items. For example, suppose that in the above scenario the test is a simple pass-fail test, and a total of 6 out of 10 is under consideration for the pass mark. Inspection of Table 1 shows that it is success on Q8 which distinguishes those with a total of 6 out of 10 from those with a total of 5 out of 10. The content of Q8 could therefore form the basis of a discussion as to whether this was indeed an appropriate cut-score.

This could allow genuine criterion referencing in standard setting. If we imagine our example is a functional maths test, and that Q8 involves calculating a percentage, if it was deemed essential that the 'minimally competent examinee' should be able to calculate a percentage, then 6 out of 10 is the lowest score on the test which guarantees this.

Also, once the standard has been set, standard-maintaining in such a scenario is also straightforward. By including a similar (ideally identical) item in a future test we might anchor the new test to the old simply by finding the lowest test total on the new test guaranteeing success on this item, assuming of course that the new test also produces scores in the Guttman pattern. Thus it would not matter if the new test were easier or more difficult than the original test – the cut-score would vary accordingly.

The traditional item analysis statistics of facility (mean item mark as a proportion of maximum item mark) and discrimination are shown below in Table 2.

Table 2: Facility values and discrimination indices for example data in Table 1

	Easy →									*Hard*
	Q4	*Q2*	*Q1*	*Q7*	*Q5*	*Q8*	*Q3*	*Q10*	*Q9*	*Q6*
Facility	0.96	0.91	0.83	0.70	0.52	0.43	0.30	0.17	0.09	0.04
Discrimination	0.34	0.44	0.57	0.68	0.77	0.78	0.72	0.59	0.45	0.34

Facility is the mean mark on each item as a proportion of maximum item mark[2].

Discrimination is the Pearson correlation between item score and total score minus that item.

These statistics do not seem especially useful for setting cut-scores at an awarding meeting. The facility values are sample-dependent, and the discrimination indices are both sample-dependent and facility-value-dependent.

It is more informative to consider the relationship between score on test and score on item. This can be presented graphically in what are known as Item Characteristic Curves (ICCs). Figure 1 shows the ICCs for Q1, Q5 and Q10 in our example.

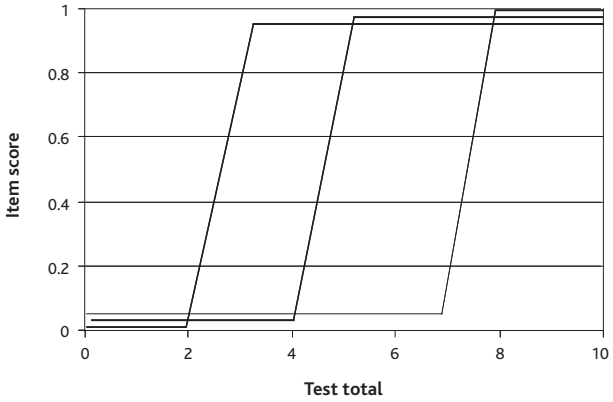


Figure 1: Plot of item score against test total for Q1, Q5 and Q10. (The top and bottom of each ICC should be at 1 and 0, but are separated here for clarity)

These ICCs illustrate the step change in performance on each item with increasing test total score. The slope of the ICC is another indicator of discrimination – in this case it can be clearly seen that each item discriminates very well (perfectly!) by the same amount at a different point on the raw score scale. This information is obscured by using the traditional discrimination statistics (see Table 2).

This kind of display shows the link between performance on the test as a whole and performance on an individual item, and as such is far more relevant to the task which awarders are engaged in when assessing performance on 'key discriminators'.

However, it is virtually impossible in practice to construct tests which produce the deterministic Guttman pattern of responses. This is first because people of the same overall ability differ in their specific knowledge and skills and thus tend to produce different patterns of correct and incorrect responses; and secondly because there are many unknown 'random' variables which might influence a particular response on a particular occasion. In order to overcome both of these factors it would be necessary to use items very widely spaced in difficulty and to administer the test to a population with a very wide distribution of ability. For example, the four-item test below, administered to the entire population of England, might produce a Guttman pattern of responses:

Q1 $2+2=?$

Q2 $^2/_3 \times {}^3/_4 = ?$

Q3 $x^2 - 5x + 6 = 0, x = ?$

Q4 Prove $e^{i\pi} + 1 = 0$

2 For a dichotomous item the facility value is also the proportion of examinees who answered correctly.

But the results of such a test would be extremely uninformative for most educational purposes! Therefore it is necessary to consider another idealisation, but a slightly more realistic one – the Rasch model.

The Rasch model

It is outside the scope of this article to derive or explain the Rasch model – see Wright & Stone (1979) or Bond & Fox (2000) for details. The Rasch model for dichotomous items can be written as:

$$p(x_{ni} = 1 | \beta_n, \delta_i) = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}$$

where x_{ni} is the score of person n on item i, β_n is the ability[3] of person n, and δ_i is the difficulty of item i.

This model can be considered to be a stochastic form of Guttman's model. This is because the pattern of expected scores from this model is the Guttman pattern. In other words, a person with higher ability has a higher probability of success on every item than a person of lower ability, and every person has a lower probability of success on a more difficult item than on an easier item.

The pattern of observed scores will not exactly conform to the Guttman pattern, but should be a stochastic approximation to it if the data fit the Rasch model. Table 3 shows some simulated data generated to fit the Rasch model, using approximate parameters[4] derived from the data in Table 1.

Table 3 shows that the rank order of examinees has changed slightly, as has the rank order of items, due to the random element in the data generation. More importantly, the data now does not exactly conform to the Guttman pattern, as can be seen by comparing the score patterns of the three examinees with a test total of 5 out of 10. E11's pattern of performance is exactly in line with expectation, but E13 and E14 have both succeeded on some more difficult items and failed some easier items.

Table 3 illustrates the problem of using 'key discriminators' for deciding on cut-scores when data does not fit the ideal Guttman pattern (which it never does). Consider the performance on Q5 by examinees with test scores of 5 and 6. Two of the three examinees with a score of 5 succeeded on this item, whereas only one of the four examinees with a score of 6 did. If consideration of this item were a main focus for awarders then 5 might seem a more appropriate cut-score than 6 – even though examinees with a score of 5 have (by definition) achieved less overall than those with a score of 6, and in particular on other items (e.g. Q8 and Q1) which might not have been deemed 'key discriminators'.

A further issue which is clarified by considering the Guttman pattern in Table 3 is that of examinee (rather than item) score profiles. Some examinees will have an 'unusual' pattern of responses in that they tend to have succeeded on items that they might have been expected to fail, and vice versa. An extreme example in Table 4 is a comparison between examinees E04 and E09 – both with a score of 1. E04 answered the easiest question correctly and failed all the others (as expected), but E09 succeeded on the second hardest question on the test. Given that awarders can only look at a small selection of the scripts on each mark, it would make sense to choose scripts from examinees whose pattern of responses conforms reasonably closely to the Guttman pattern. This is because their responses best exemplify what the test was measuring. In the real world the patterns are far more 'messy' than the neat example in Table 3, which was generated to fit the Rasch model, but the principle is still relevant.

Table 3: Pattern of scores generated by the Rasch model

	Easy →									*Hard*	
	Q2	*Q4*	*Q1*	*Q7*	*Q8*	*Q3*	*Q5*	*Q10*	*Q9*	*Q6*	*Total*
E01	0	0	0	0	0	0	0	0	0	0	0
E04	1	0	0	0	0	0	0	0	0	0	1
E09	0	0	0	0	0	0	0	0	1	0	1
E03	1	1	0	0	0	0	0	0	0	0	2
E05	1	1	1	0	0	0	0	0	0	0	3
E06	0	1	1	1	0	0	0	0	0	0	3
E07	1	1	1	0	0	0	0	0	0	0	3
E02	1	1	1	0	1	0	0	0	0	0	4
E10	1	1	1	1	0	0	0	0	0	0	4
E12	1	1	1	1	0	0	0	0	0	0	4
E11	1	1	1	1	1	0	0	0	0	0	5
E13	1	1	0	1	0	1	1	0	0	0	5
E14	1	1	0	1	1	0	1	0	0	0	5
E08	1	1	0	1	0	1	1	1	0	0	6
E15	1	1	1	1	1	1	0	0	0	0	6
E16	1	1	1	1	1	0	0	0	1	0	6
E17	1	1	1	1	1	1	0	0	0	0	6
E18	1	1	1	1	1	1	1	0	0	0	7
E20	1	1	1	1	1	1	1	0	0	0	7
E19	1	1	1	1	1	0	1	1	1	0	8
E21	1	1	1	1	1	1	1	1	0	0	8
E22	1	1	1	1	1	1	1	1	1	1	10
E23	1	1	1	1	1	1	1	1	1	1	10

3 Ability here does not mean some innate ability or IQ – it simply means the examinee's level on the trait presumed to underlie performance on the test.

4 Note for Rasch experts: it is impossible to estimate parameters for data which exactly fit a Guttman pattern, hence the 'approximate'. The logit difficulties were derived by transforming the facility values, then the abilities were estimated iteratively, arbitrarily assigning reasonable values to scores of 0 and 10 respectively.

Practical application

For an exam with a large entry (say >1000) we can calculate the average score on each item for the set of examinees with each possible score on the test. A table would be one way to present this information for awarders. There should be a general increase in score on each question as test total score increases – this is more likely to be the case for a question which discriminates well (by definition) and the increase is likely to be smoother when the number of examinees is large. At very high and very low test total scores there is likely to be some fluctuation because of the low numbers of examinees.

Table 4 illustrates this kind of information for a GCSE paper, where ILD from approximately 38,000 examinees was captured. Note that the information is shown at the level of the whole question. It would also be possible to show the same information at sub-question level, but such a table would potentially be very large, creating a danger of 'information overload'.

The information in Table 4 might be easier to appreciate if it were presented in graphical form.

The graphs in Figure 2 show one possibility for creating visually informative displays. They simply join the mean y-values (score on question) for each value of x (score on test), and show ± 2 standard errors of each mean. This conveys the information that the location of the mean is more variable at the extremes (or wherever N is low), and also takes into account the spread of the y-values at each value of x (the formula for the standard error is $\sigma/\sqrt{N}$, where s is the standard deviation of the y-values). The individual data points are not shown in the graphs in

Figure 2. This makes the graphs less cluttered, but leaving the points on would emphasise the extent to which there is variability at the individual question level for a given score on the test as a whole.

With a smaller cohort it might be preferable to fit a smoothed line through all the data points, rather than joining up the means in a 'dot-to-dot' fashion. This is an area for further practical experimentation.

Use of item level information in an award meeting

How might the information shown in Table 4 and Figure 2 be used in an awarding meeting? First of all we should note that these whole questions vary quite a lot in terms of maximum marks (Q5 is out of 3 marks and Q9 is out of 12 marks). Q7 was clearly too difficult for most examinees. Q1 and Q8 discriminated best for examinees at the lower end of the score range. The questions with larger mark totals discriminated more smoothly across the score range, as might be expected.

It is possible to identify two approaches for linking information about examinees' performance on individual questions with grading decisions – what we might call a 'prescriptive' approach and a 'maintaining' approach.

On the prescriptive approach, expert judgement combined with grade descriptors might be used to make pronouncements like 'The average borderline grade C examinee ought to score at least 7 marks on Q3'. From the ICC for Q3 or from Table 4, this can be seen to imply a cut-score of around 41. Making several pronouncements of this type on different questions from different topic areas across the paper would produce several potential cut-scores – these could then form a purely judgementally derived range of marks to consider for the grade C boundary. This approach might be more effective at sub-question level because more information could be used (but this would have to be balanced against the dangers of information overload).

On the maintaining approach (which can only work when ILD from two or more sessions are available) the awarding panel would identify questions on the current paper which are similar enough to questions on a previous paper for it to be reasonable to expect performance on them to be equivalent. Now the argument would be along the following lines: 'Last year the borderline grade C examinees (with a test score of 40) averaged 1.2 out of 2 on question 7a, which required them to label a diagram of a cell. This year's question 3b was practically identical, and examinees who averaged 1.2 out of 2 scored 42 on the test overall, suggesting a mark of 42 would be appropriate for this year's boundary.'

Obviously the more 'equivalent' questions that can be identified, the better the linking will be (and this need not always link back to the previous session – links which go back further will help avoid 'drift' in boundaries). There are obviously many caveats which could be raised, such as the extent to which the questions really are equivalent[5], changes over time in topic relevance, drifts in item difficulty, teaching trends etc. – a microcosm of the debates around standards over time more generally! But it is nonetheless a method with some rational justification.

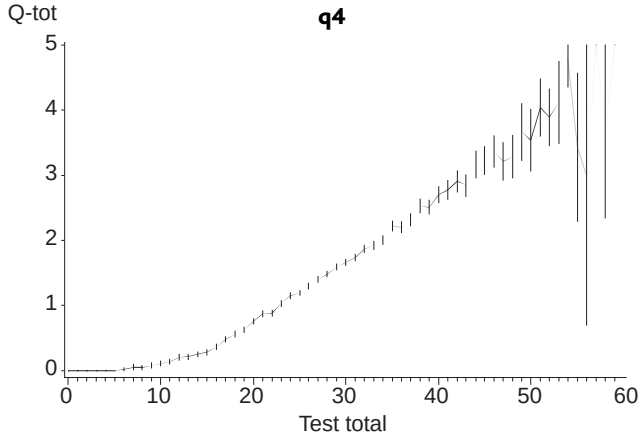
Because of the wide variability in question performance across individual scripts, these judgements based on the ICCs (which show the average performance of the entire examination cohort) might be found to be more effective than judgements based on scrutinising a tiny proportion of the scripts on each mark (as currently happens). It would also prevent individual awarders' judgements being skewed by their impressions from having marked a possibly unrepresentative batch of scripts. For example, using the above data we can see from the graph for Q1 that this question was discriminating most effectively for pupils with a test score of between about 10 and 25 marks. Repeatedly sampling two scripts at random from those with a test total 4 marks apart, and then 1 mark apart gave the results in Table 5 below.

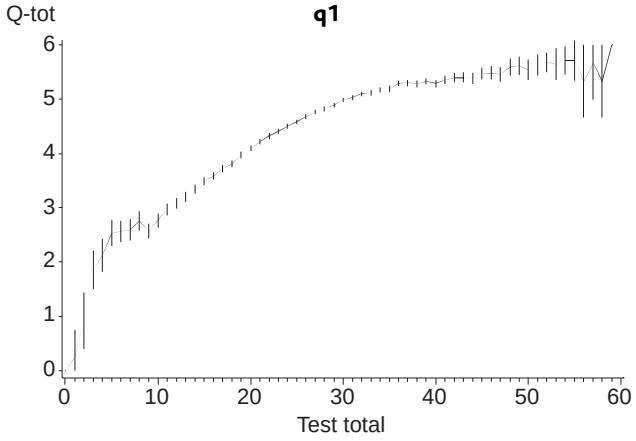
5 It has sometimes been observed that small changes to questions can have a large effect on their facility value, so judgements of equivalence should be made with great care.

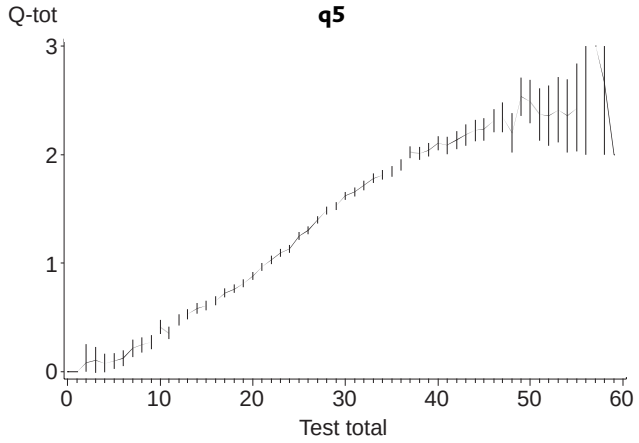
Table 4: GCSE paper – mean scores on each question for pupils with each total test score

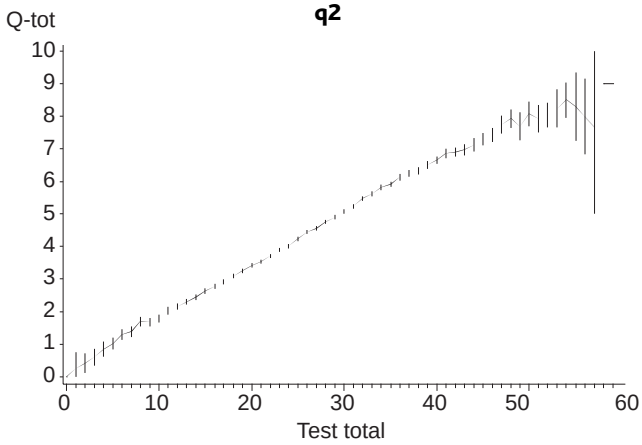
	Max mark→	*6*	*10*	*11*	*5*	*3*	*4*	*4*	*4*	*12*	*7*
Test total	*N*	*q1*	*q2*	*q3*	*q4*	*q5*	*q6*	*q7*	*q8*	*q9*	*q10*
0	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	4	0.25	0.25	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00
2	12	0.92	0.42	0.00	0.00	0.08	0.33	0.00	0.17	0.08	0.00
3	28	1.86	0.61	0.04	0.00	0.11	0.14	0.00	0.14	0.07	0.04
4	39	2.13	0.85	0.13	0.00	0.08	0.15	0.00	0.13	0.36	0.18
5	71	2.54	1.01	0.48	0.00	0.10	0.28	0.00	0.17	0.30	0.13
6	99	2.57	1.29	0.65	0.02	0.12	0.34	0.00	0.31	0.45	0.24
7	117	2.60	1.38	1.03	0.05	0.21	0.49	0.00	0.45	0.53	0.26
8	169	2.76	1.70	1.21	0.05	0.25	0.66	0.01	0.57	0.50	0.30
9	211	2.57	1.68	1.65	0.08	0.27	0.63	0.00	0.82	0.82	0.47
10	263	2.77	1.79	1.83	0.11	0.41	0.79	0.00	1.04	0.76	0.51
11	338	2.97	2.02	1.93	0.14	0.36	0.83	0.01	1.26	0.88	0.62
12	458	3.08	2.15	2.13	0.21	0.48	0.90	0.01	1.29	0.99	0.76
13	579	3.20	2.30	2.32	0.21	0.53	0.88	0.02	1.54	1.19	0.83
14	643	3.35	2.45	2.40	0.25	0.58	1.02	0.01	1.73	1.28	0.94
15	752	3.49	2.63	2.60	0.28	0.61	1.05	0.03	1.82	1.44	1.05
16	925	3.58	2.77	2.68	0.37	0.65	1.12	0.04	2.02	1.63	1.14
17	992	3.72	2.92	2.76	0.48	0.72	1.12	0.03	2.19	1.81	1.25
18	1139	3.81	3.09	2.88	0.56	0.76	1.19	0.04	2.32	2.02	1.33
19	1227	3.97	3.26	2.97	0.63	0.81	1.22	0.04	2.46	2.20	1.44
20	1437	4.09	3.42	3.06	0.76	0.88	1.32	0.06	2.51	2.37	1.53
21	1542	4.22	3.53	3.14	0.87	0.96	1.35	0.06	2.68	2.57	1.61
22	1470	4.32	3.71	3.21	0.88	1.03	1.36	0.06	2.81	2.89	1.73
23	1708	4.41	3.88	3.33	1.03	1.09	1.38	0.08	2.88	3.08	1.83
24	1808	4.50	4.01	3.48	1.15	1.13	1.46	0.08	2.96	3.33	1.90
25	1780	4.58	4.23	3.53	1.19	1.25	1.54	0.09	3.04	3.53	2.02
26	1856	4.68	4.43	3.69	1.29	1.30	1.60	0.10	3.08	3.71	2.12
27	1907	4.77	4.55	3.80	1.40	1.40	1.64	0.11	3.16	3.95	2.22
28	1788	4.82	4.75	3.90	1.49	1.49	1.69	0.11	3.21	4.21	2.33
29	1687	4.89	4.90	4.09	1.59	1.53	1.70	0.12	3.31	4.44	2.44
30	1656	4.99	5.07	4.26	1.66	1.62	1.81	0.13	3.31	4.62	2.54
31	1519	5.03	5.24	4.40	1.73	1.66	1.88	0.13	3.36	4.88	2.69
32	1418	5.10	5.48	4.59	1.86	1.71	1.88	0.16	3.36	5.05	2.81
33	1250	5.12	5.62	4.84	1.92	1.78	1.99	0.12	3.40	5.31	2.89
34	1111	5.17	5.82	5.09	2.00	1.81	2.03	0.15	3.42	5.56	2.95
35	1011	5.19	5.91	5.28	2.21	1.85	2.09	0.15	3.50	5.75	3.06
36	843	5.29	6.13	5.58	2.20	1.90	2.12	0.17	3.49	5.94	3.18
37	766	5.30	6.25	5.68	2.31	2.02	2.24	0.16	3.52	6.20	3.31
38	606	5.29	6.32	6.15	2.53	2.01	2.19	0.18	3.54	6.39	3.41
39	548	5.33	6.51	6.28	2.51	2.04	2.32	0.20	3.52	6.74	3.55
40	497	5.29	6.65	6.58	2.70	2.10	2.42	0.18	3.51	6.95	3.61
41	348	5.36	6.86	6.99	2.77	2.08	2.31	0.20	3.66	7.12	3.65
42	314	5.40	6.90	7.28	2.91	2.13	2.50	0.22	3.53	7.39	3.74
43	236	5.40	6.97	7.60	2.84	2.18	2.54	0.27	3.62	7.59	3.98
44	182	5.38	7.12	7.75	3.16	2.22	2.57	0.27	3.62	7.88	4.03
45	160	5.47	7.30	7.69	3.23	2.23	2.73	0.29	3.71	8.13	4.23
46	136	5.49	7.43	7.99	3.36	2.31	2.78	0.27	3.59	8.43	4.36
47	102	5.46	7.74	8.60	3.22	2.34	2.73	0.48	3.64	8.65	4.16
48	70	5.59	7.93	8.63	3.29	2.20	3.06	0.51	3.66	8.61	4.53
49	45	5.62	7.69	8.47	3.67	2.53	2.91	0.56	3.76	8.89	4.91
50	37	5.54	8.08	8.86	3.54	2.49	3.00	0.89	3.68	9.24	4.68
51	27	5.63	7.93	9.11	4.04	2.37	3.11	0.85	3.67	9.59	4.70
52	28	5.68	8.04	9.43	3.89	2.36	3.00	0.71	3.71	10.04	5.14
53	17	5.65	8.24	9.65	4.12	2.41	3.41	0.88	3.94	9.65	5.06
54	14	5.71	8.50	9.71	4.79	2.36	2.93	0.86	3.86	10.00	5.29
55	7	5.71	8.29	10.43	3.43	2.43	3.29	1.43	3.86	10.29	5.86
56	3	5.33	8.00	10.33	3.00	2.67	4.00	1.33	4.00	11.33	6.00
57	3	5.67	7.67	9.67	5.00	3.00	2.67	1.67	4.00	10.67	7.00
58	3	5.33	9.00	10.00	3.67	2.67	4.00	1.67	3.67	11.33	6.67
59	1	6.00	9.00	10.00	5.00	2.00	4.00	1.00	4.00	12.00	6.00

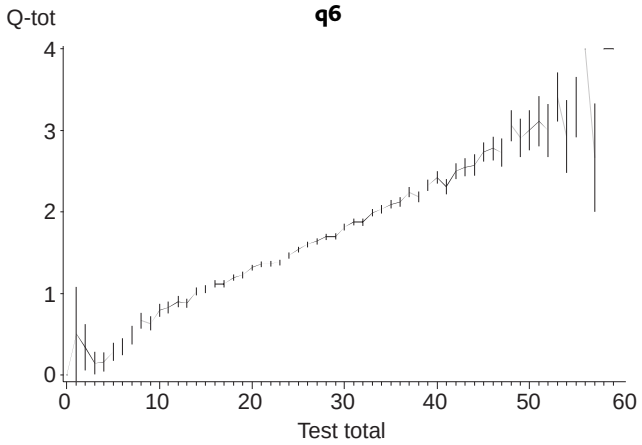
Figure 2: Question score v test score for each question on the GCSE paper

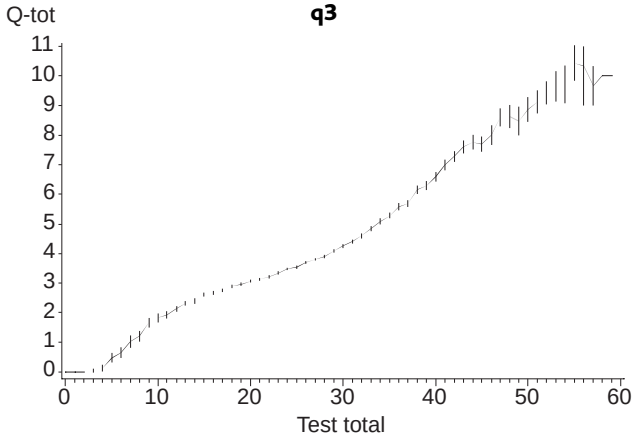


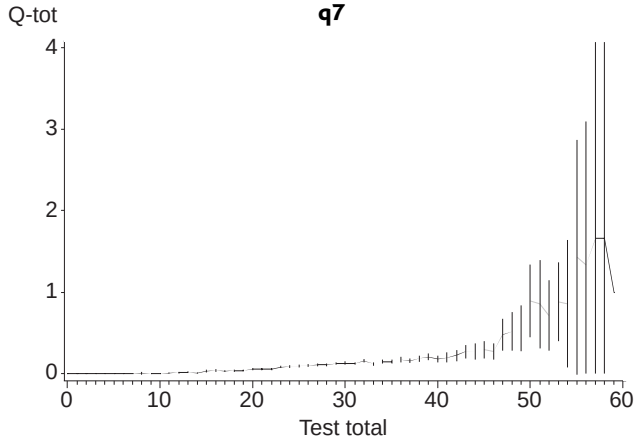


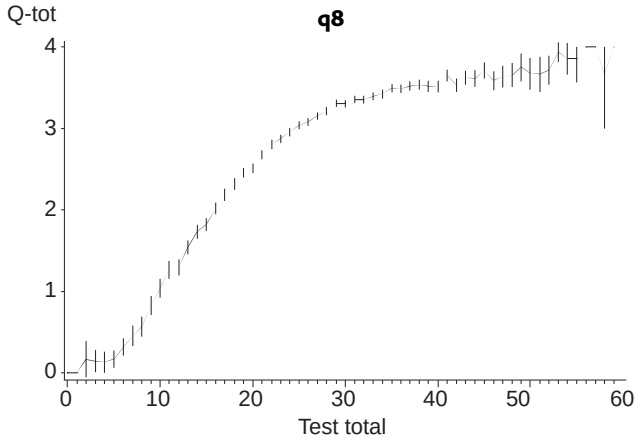


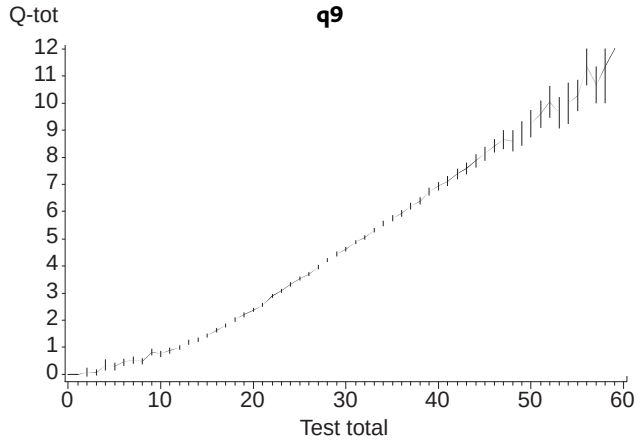












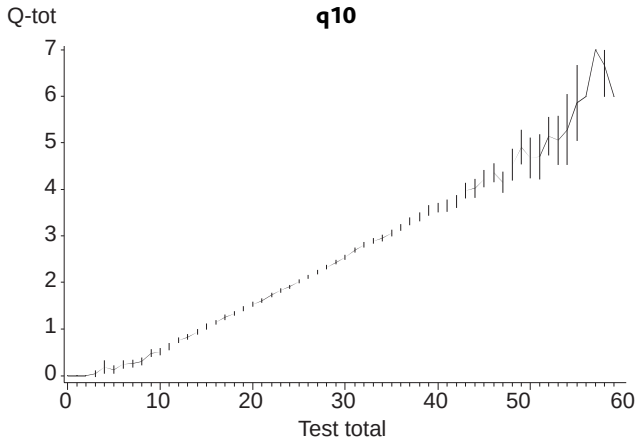


Table 5 shows that even in the most discriminating part of the range, with scripts 4 marks apart on total score the script with the higher test total only scored a higher mark on Q1 about half the time, dropping to around 40% of the time with scripts 1 mark apart. This suggests that script scrutiny might not be a good way to relate performance on 'key discriminators' to total test score.

This is not to suggest that the use of ICC graphs means that script scrutiny can be dispensed with altogether. It does, however, suggest a different focus for the script scrutiny. The ICC graphs can be used to identify the 'key discriminators' for a particular boundary and to derive expectations about the likely range of test total marks corresponding to that boundary, either using the 'prescriptive' or the 'maintaining'

Table 5: Result of 10,000 comparisons of pairs of scripts sampled at random

Performance on Q1	*Test totals 4 marks apart (18 and 22)*	*Test totals 1 mark apart (19 and 20)*
Script with higher test total better	51%	39%
Scores equal	26%	29%
Script with lower test total better	23%	22%

approach described above. The role of the script scrutiny would then be to make a global, holistic judgement about examinee performance on scripts in that range, taking into account performance on all the questions. We must not forget that examinees can compensate for poor performance on the key discriminators by good performance elsewhere. The judgemental task could now perhaps be phrased along the lines 'Would you be happy for scripts in this mark range to receive a C grade?'

In summary, the approach might work as follows:

1. The awarding panel decides for which (if any) questions a 'prescriptive' approach is appropriate and for which questions (if any) a 'maintaining' approach is appropriate.
2. For the 'prescriptive' questions, the awarding panel decides what the minimum mean mark on those questions for examinees at a particular grade boundary should be, using expert judgement and (if appropriate) grade descriptors. The test total mark corresponding to this mark is then located using the ICCs.
3. For the 'maintaining' questions, the awarding panel uses the ICCs to locate the test total mark corresponding to the same question mean mark as that obtained by borderline examinees in a previous session.
4. Steps 2 and 3 should now have created a range of test total marks for consideration at each judgemental boundary. Each range can now be compared with the range produced at the pre-award based on statistical information about score distribution and cohort composition. Hopefully, there will be some overlap between these ranges!
5. The awarding panel can scrutinise scripts in the overlapping range to ratify a particular mark, or narrower range, as appropriate for the boundary in question.
6. The final boundary mark is agreed by the usual process of considering all available evidence.

Implementing this kind of process would create a system where the judgements about scripts can be less influenced by information about pass rates and cohort composition. It has been argued elsewhere (e.g. Black & Bramley, 2008) that this would be desirable.

Conclusion

In summary, consideration of the idealised Guttman pattern of examinee scores on test items leads to the following conclusions:

- If only a small number of scripts is chosen for scrutiny at an award meeting, it is possible that performance on an item designated as a 'key discriminator' will not correspond well with the total score.
- Traditional item analysis statistics (facility values and discriminations) may not be particularly useful for identifying the 'key discriminators' at each grade boundary, but empirical Item

Characteristic Curves (ICCs), ideally based on points plotted at each possible score on the test (when enough data are available), could be much more useful.

- Scripts for the award meeting could be screened to eliminate 'misfitting' examinees with unusual response patterns, or positively selected to aim for responses which conform as well as possible to the Guttman pattern.
- An explicit rationale should be provided for how the item level data will be used in making decisions about grade boundaries – for example, the 'prescriptive' and 'maintaining' rationales described in this article.

References

Black, B. & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education*, **23**, 3, 357–373.

Bond, T. & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B. D. & Stone, M. (1979). *Best Test Design*. Chicago: MESA Press.

Statistics Reports Series

The Statistics Team Research Division

The ongoing 'Statistics Reports Series' provides statistical summaries of various aspects of the English examination system such as trends in pupil attainment, qualifications choice, combinations of subjects and subject provision at school. These reports, produced using national-level examination data, are available in .pdf format on the Cambridge Assessment website: http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Statistical_Reports

In 2009, the following reports were produced:

- Statistics Report Series No.8: Uptake of GCSE AS level subjects in England, 2001–2007
- Statistics Report Series No.9: Numbers achieving 3 A grades in specific A-level combinations by school type and LEA
- Statistics Report Series No.10: Some issues on the uptake of Modern Foreign Languages at GCSE
- Statistics Report Series No.11: Uptake of GCSE and A-level subjects in England by Ethnic Group, 2007
- Statistics Report Series No.12: A-level uptake and results by gender, 2002–2007
- Statistics Report Series No.13: GCSE uptake and results by gender, 2002–2007

Other statistical reports also available on the Cambridge Assessment website are:

- Statistics Report Series No.1: Provision of GCE A-level subjects
- Statistics Report Series No.2: Provision of GCSE subjects
- Statistics Report Series No.3: Uptake of GCE A-level subjects in England, 2001–2005
- Statistics Report Series No.4: Uptake of GCSE subjects, 2000–2006
- Statistics Report Series No.5: Uptake of GCE A-level subjects in England, 2006
- Statistics Report Series No.6: Numbers of A-level examinations taken by candidates in England 2006 and the percentages attaining 3 or more A grades
- Statistics Report Series No.7: The relationship between A-level grade and GCSE grade by subject

Factsheets

The Statistics Team Research Division

In order to make our research accessible to a wider audience we have produced a series of easy-to-read factsheets. The objective of these factsheets is to 'headline' the main findings of some research projects.

They are available in .pdf format on the Cambridge Assessment website:http://www.cambridgeassessment.org.uk/ca/Our_Services/Research

The full research reports can be found in the 'Conference Papers' section of the same website.

As of December 2009, factsheets on the following subjects have been produced:

- AS and A-level choice: Ten factsheets based on the project entitled 'A-level subject choice in England: Patterns of uptake and factors affecting subject references'.
- Emotional Intelligence: Three factsheets based on the project entitled 'Can trait Emotional Intelligence predict differences in attainment and progress in secondary school?'

Research News

Cambridge Assessment Network

4th Cambridge Assessment Conference

On 19 October 2009 around 150 education and assessment professionals gathered at Robinson College for the 4th Cambridge Assessment Conference. Taking the theme *'Issues of control and innovation: the role of the state in assessment systems'*, the conference encouraged debate on key issues at a significant time. Major changes to regulation, the shape of agencies, and to the allocation and form of responsibilities are underway, and mapping the consequences and implications of these changes is a vital process.

The conference featured keynote speeches from Professor Robin Alexander, Director of the Cambridge Primary Review, and Professor Alison Wolf, King's College London. In addition delegates were able to attend three from a selection of nine discussion seminars, given by renowned experts such as Professor Mary James (University of Cambridge), Isabel Nisbet (Ofqual), and Dr John Allan (SQA). During the seminar sessions delegates had the opportunity to comment on and debate the issues, and ask questions of the speakers.

A drinks reception was held at the close of the conference giving delegates further opportunity to network and discuss the issues with like-minded professionals.

Paul Newton, Director of the Cambridge Assessment Network commented: "Looking back over the years, the number of changes to education and assessment policy made by the state is overwhelming. With the establishment of Ofqual, continued crises within the UK assessment system and the ongoing discussion of standards, Cambridge Assessment is pleased to have provided a platform for debating the crucial issue of how far the state should be controlling educational assessment, with a conference programme that included such prestigious speakers."

Presentations and audio clips from the conference can be found at http://www.cambridgeassessment.org.uk/ca/Viewpoints/Viewpoint?id=131163.

Forthcoming Network Events

Assessing Citizenship in schools. Are we measuring the unmeasurable?

On 11 March Dr Mary Richardson from the Centre for Beliefs, Rights and Values in Education, Roehampton University, will be coming to talk about assessing Citizenship in schools. The seminar discusses the problems of assessing Citizenship, a subject which is perceived by many teachers as unconventional, and by some, as unassessable. The value of assessing Citizenship will be considered in the context of the findings from an empirical study conducted in secondary schools across England. The challenge for Citizenship educators is the need for meaningful assessments that relate to the beliefs and values under discussion in lessons.

When is an exam not really an exam?

On 17 March Dr Sue Horner will be presenting one of the Network's Forum seminars. The effectiveness of techniques to assess learner progress and performance is related to the purposes for which assessment is undertaken. The roles of different techniques in formal and informal assessment need further exploration. When roles and purposes have high stakes associated with them this impacts the selection of techniques. Public confidence is often cited as a reason for resisting a range of styles of assessment. Are there ways forward for teacher assessment, tests and qualifications?

For further details of these events, or to receive a monthly update of forthcoming events, or a copy of the Network's Programme for 2010, please contact the Network Team at: thenetwork@cambridgeassessment.org.uk.

Full details of our events programme are available at: www.assessnet.org.uk.

Conferences and seminars

British Educational Research Association

The BERA Annual conference was held from 2–5 September 2009 at the University of Manchester. Colleagues from the Research Division and CIE presented the following papers:

Victoria Crisp: *Objective questions in GCSE science: Exploring question difficulty, item functioning and the effect of reading difficulties*.

Milja Curcin, Beth Black and Tom Bramley: *Standard-maintaining by expert judgement: Using the rank-ordering method for determining the pass mark on multiple-choice tests*.

Tim Gill, Carmen Vidal Rodeiro and John F. Bell: *Aspects of AS and A-level Physics uptake*.

Jackie Greatorex: *How are archive scripts used in judgements about maintaining grading standards?*

Martin Johnson, Rita Nádas and Hannah Shiell: *An investigation into marker reliability and other qualitative aspects of on-screen marking*.

Stuart Shaw and Victoria Crisp: *What was this student doing? Evidencing validity in A-level assessments*.

Nicholas Raikes, Jane Fidler and Tim Gill: *Must examiners meet in order to standardise their markings? An experiment with new and experienced examiners of GCE AS Psychology*.

Full details of the papers can be found on the Cambridge Assessment websitehttp://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Conference_Papers

European Association for Research on Learning and Instruction

In August Rita Nádas attended the EARLI conference in Amsterdam and presented research on: *Markers' metacognition: Does metacognitive*

intervention enhance marking performance and cognitive marking strategy usage?

International Association of Educational Assessment

The 35th Annual IAEA conference took place in Brisbane, Australia from 13–18 September. The theme of the conference was 'Assessment for a Creative World'. Colleagues from Cambridge Assessment presented the following papers:

Newman Burdett and Martin Johnson: *Intention, interpretation and implementation: Some paradoxes of assessment for learning across educational contexts*.

Stuart Shaw and Martin Johnson: *Annotating essays on screen: The influence of reading environment on annotative practice and assessor comprehension building*.

Martin Johnson, Rita Nádas and Sylvia Green: *Marking essays on screen: an investigation into the reliability of marking extended subjective texts*.

European Conference on Educational Research

In September Irenka Suto attended the ECER conference in Vienna and presented a paper on: *How should grade boundaries be determined in examinations? An exploration of the script features that influence expert judgements*.

Association for Educational Assessment – Europe

The theme of the 10th Annual Conference of AEA-Europe, which took place in Malta from 5–7 November, was: 'Innovation in Assessment to meet changing needs'. Papers by the following researchers were presented:

Martin Johnson, Rita Nádas and Sylvia Green: *Marking essays on screen: an investigation into the reliability of marking extended subjective texts*.

Tom Bramley: *The effect of manipulating features of examinees' scripts on their perceived quality*.

Beth Black: *Introducing a new subject and its assessment in schools: The challenges of introducing Critical Thinking AS/A level in the UK*.

Stuart Shaw, Victoria Crisp and Nat Johnson: *A proposed framework for evidencing assessment validity in large-scale, high-stakes international examinations*.

Tim Oates and Jill Grimshaw: *How can we help teachers respond to national assessment strategies? The position in England*.

Four colleagues have received professional recognition from the Association – Paul Newton and Newman Burdett as Fellows, and Jill Grimshaw and Steve Murray as Practitioners.

Irenka Suto, Stuart Shaw and Jo Ireland won the best poster prize for their poster on *Creating research programmes to support the development and validation of qualifications: What are the key assessment issues and what are the key research methods?*

Publications

The following articles have been published since Issue 8 of *Research Matters*:

Crisp, V. & Novaković, N. (2009). Is this year's exam as demanding as last year's? Using a pilot method to evaluate the consistency of examination demands over time. *Evaluation and Research in Education*, **22**, 1, 3–15.

Emery, J.L. & Bell, J.F. (2009). The predictive validity of the BioMedical Admissions Test for pre-clinical examination performance. *Medical Education*, **43**, 6, 557–564.

Green, S. & Oates, T. (2009). Considering alternatives to national assessments in England: possibilities and opportunities. In: C. Whetton (Ed.), National Curriculum Assessment in England: how well has it worked? Perspectives from the UK, Europe and beyond. *Educational Research Journal, Special Issue*, **51**, 2, 229–245.

For all the latest research news please visit http://www.cambridgeassessment.org.uk/ca/Our_Services/Research

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

Tel: 01223 553854
Fax: 01223 552700
Email: ResearchProgrammes@cambridgeassessment.org.uk
http://www.cambridgeassessment.org.uk

ISSN: 1755–6031