Centre for Education Research and Policy

Testing the validity of judgements about geography essays using the Adaptive Comparative Judgement method

Claire Whitehouse

Abstract

Adaptive comparative judgement (ACJ) is an alternative to marking that presents judges with pairs of students' work and asks them to decide, holistically, which piece of work contains more of a specified trait or set of traits. There are a number of reports on the highly reliable rank orders achieved using ACJ. However, none of these reports addresses the validity of the criteria on which judges base their decisions.

The reliability of the rank ordering of 564 AS-level geography essays by 23 teachers or examiners of geography was reported previously. The judges in this study were asked to use their professional judgement when making decisions about essays; they were not provided with mark schemes or assessment objectives, but two importance statements were made available to them. After each judgement (92.4% of the total), the judges in this empirical study made notes about what was to the forefront of their minds when they made a decision between two essays. The investigation reported here uses thematic analysis of these notes to identify and test the validity of the criteria the judges used to make their decisions.

On the whole the judges used the language of the mark scheme and the assessment objectives when describing the knowledge demonstrated in the essays. They used language from these two sources to a lesser extent when describing skills, but nonetheless indirect links could be established between the content of the notes and existing documents. These links demonstrate the use of existing shared criteria by the judges, thus the validity of the criteria used by judges in their decision-making was confirmed. However, these criteria are already established as part of examiner training, marking and teacher support. This has implications for the introduction of ACJ as a replacement for marking, which are discussed.

Key words: adaptive comparative judgement, psychological validity, paired comparisons, shared criteria

Introduction

Purpose

A recent empirical study investigated the feasibility of using adaptive comparative judgement (ACJ) as an alternative to marking in national high stakes assessments (Whitehouse & Pollitt, 2012). One of the outcomes of the study was to place 564 essay responses to an AS-level geography paper in a rank order with a reliability of 0.97, which is high when compared with the reliabilities achieved in traditional marking. The rank order was based on quality parameters that were estimated using the Rasch model to fit the decisions from a number of paired comparisons of the essays. The paired comparisons were made by 23 judges who were teachers or examiners of geography.

AQA Education (AQA) is a registered charity (number 1073334) and a company limited by guarantee registered in England and Wales (number 3644723). Our registered address is AQA, Devas Street, Manchester M15 6EX.

One of the judges' tasks in the study was to make a note after each decision as to what they were thinking about when they made their decision. It is the analysis of the content of these notes that is reported here as a way of testing the validity of the judgements. The validity under test is that referred to by Bramley (2007) as the "psychological validity" or whether it is possible for judges making paired comparisons of essays to observe sufficient and appropriate construct-relevant features on which to base their judgements.

The aim of the current investigation was to explore the following research questions.

- What features of the essays are influencing the judges' decisions in adaptive comparative judgement?
- Is it possible to relate the source or sources of these influences to existing criteria?

Paired comparisons and adaptive comparative judgement (ACJ)

The paired comparisons method was first proposed by Thurstone (Thurstone, 1927a, b) to investigate a range of psychological phenomena involving judgement. It was first used in the cross-moderation strand of comparability studies in educational assessment in 1996 to compare some new modular A level syllabuses with their legacy linear equivalents (D'Arcy, 1997). Since then it has been used by the qualifications regulators (QCA, 2006) and awarding organisations in the UK to monitor performance standards over time and between awarding organisations (see, for example, Adams and Pinot de Moira (2000), Edwards and Adams (2002) and Fearnley (1999).

In a study using paired comparisons, judges, working independently, are presented with pairs of objects. In an assessment context the objects are usually students' scripts from examinations and either coursework or controlled assessment. However, paired comparisons can take place at the level of individual questions and responses, such as essays. So, for each pair of essays presented to them, judges are asked to decide which essay shows more of a particular trait. The outcomes from these decisions are used to construct a common scale of 'perceived quality' on which all essays are placed. An essay's location on the scale is dependent on the number of times it 'wins' or 'loses' the comparisons it takes part in and the locations on the scale of the essays to which it is compared.

The method of paired comparisons has been put forward as an alternative to marking. The advantage of the paired comparisons method when used in educational assessment is that it removes judges' individual tendencies towards leniency or severity. In comparative judgement it is the relative performance of the essays that matters, not how good, in absolute terms, a judge may think each essay is. Using a latent trait model, such as the Rasch model (Andrich, 1978), to analyse the outcomes of judges' decisions confers two additional advantages. First, the model is able to handle missing data so a full design in which every judge compares every possible pairing of essays is unnecessary (Pollitt, 2012). Second, by fitting the outcomes of decisions to a predictive model, the differences between expected and observed outcomes can be used to identify problematic scripts and inconsistent judges. Thus, quality control is available under operational conditions.

Logistical problems have impeded the wide spread use of paired comparisons in summative assessment. Not the least of these is the provision to the judges of the materials to be compared and the completion of the analysis in an appropriate time frame (Bramley, 2007; Fowles, 2000; Jones, 1997). *Adaptive* comparative judgement (ACJ) overcomes some of these problems by exploiting improvements in computer processing speeds and connectivity (Pollitt, 2012). ACJ allows judges to carry out paired comparisons remotely as the pairs of essays are delivered using a web-based browser. The time lag between making decisions about the paired comparisons and the statistical analysis of the decisions is greatly reduced. This allows subsequent pairs for comparison to be selected for efficiency in gaining information.

Studies that have used (adaptive) comparative judgement in educational assessment settings have focused on the high reliability of the achieved rank order (see, for example, Kimbell *et al.*, 2009; Newhouse, 2011; Pollitt, 2012; Whitehouse & Pollitt, 2012). They emphasised that the method of comparative judgement harnesses the professional expertise of large numbers of teachers as judges of the quality of student work. Nevertheless, none of these studies has attempted to address the validity of the criteria used by the judges to make their judgements.

Criteria for making judgements

As Brooks (2009) observed in the context of marking "Judgment does not take place in a vacuum; it requires some form of comparator." With the comparison of scripts against mark schemes in marking and the comparison of live scripts with archive scripts and grade descriptors in grading, it is easy to see what the comparators are. The comparator sets the standard against which the script is judged. But what guides the decision-making of judges comparing one script against another?

As an alternative to marking, judges working with paired comparisons assess holistically, if only for reasons of speed. Therefore in ACJ judges need to be familiar with a set of shared criteria on which to base their judgements. Shared criteria take many forms, for example, mark schemes, assessment objectives, grade descriptors, importance statements and specifications, all contribute to the shared pool of information about an assessment. Part of a teacher's or examiner's professional expertise is based on their working knowledge of such criteria and how to apply them. A recent study provided evidence that, without additional training, judges need to be teaching at the same level as the qualification to be able to make consistent judgements about AS-level geography essays (Whitehouse & Pollitt, 2012).

In a regulatory climate of transparency it is considered to be good practice to publish the criteria used to assess performance. A corollary to this is that it is good practice to link the marks awarded to students' work back to the published criteria. This linking back tends not to be done when judgement is holistic, making it impossible to be certain which features of a script were rewarded. From a research perspective it means that the script features that influence judgement, whether relevant or irrelevant to the construct, remain hidden.

Identifying criteria for making judgements

Two methods are most frequently used to investigate the features of assessments that might influence markers' or graders' decisions: (i) an adaptation of the Kelly's Repertory Grid (KRG) technique and, (ii) the think aloud method. Both of these methods tend to be used for the qualitative element in designed studies.

The adapted KRG technique has been used in comparability studies to compare the demand of assessments (see, for example, Adams & Pinot de Moira, 2000; Fearnley, 1999) and Johnson and Nádas (2012) reviewed its use in such studies. Used in its original context of personal construct theory, the constructs elicited from individuals using KRG are well known to the individual and are idiosyncratic. The demands of assessments, on the other hand, are complex, abstract phenomena that are difficult to work with and, although they should, may not have the same meaning for all participants (Johnson & Nádas, 2012). Ideally, all judges should generate identical constructs and rate them identically. Therefore, careful selection of judges who share a set of criteria for making their judgements is crucial, otherwise the validity of the comparability study is called into question.

The think aloud method has been used to investigate the features of scripts that markers (Crisp, 2010; Greatorex, 2008) and graders (Greatorex, 2002) pay attention to and that may influence them in their decision-making. The method asks participants to verbalise their thoughts about the information they are accessing and using whilst they undertake a task. Proponents of the method claim that the cognitive processes associated with the task are elicited concurrently with the carrying out of the task. It is argued that as a participant's conscious effort is focused on the task there is no room for reflection and verbalisation does not interfere with the task (Ericsson & Simon, 1998; van Someren, Barnard, & Sandberg, 1994, pg.26). However, Pashler 's (1994) research into how working on more than one mental task at a time reduces effectiveness contradicts this. He suggests that working memory is either devoted to types of task or is limited and needs to be re-directed when more than one task is being carried out. In their review of the use of verbal reports to access participants' cognitive processes Nisbett and Wilson (1977) also offer a challenge to the idea that participants are truly aware of the full range of their own cognitive processes and how they respond to stimuli.

In an attempt to bring statistical rigour to the study of how judges make decisions about scripts, Suto and Novaković (2012) prepared an inventory of nine script features based on a review of the literature on the features of examination scripts that influence judges who undertake marking and grading. They asked thirty judges from each subject to rate scripts from GCE biology and GCSE English using this inventory. As with KRG and the think aloud method it was not possible to know that each judge gave the same meaning to each script feature in the inventory. Nor, given the remote working, was it possible to know how much time was spent considering each script and then considering the ratings of features of that script.

Methodology

Selected method

A pure think aloud method was unsuitable in the current study as the judges worked remotely and at speed to simulate the use of ACJ under live assessment conditions. This approach conserved environmental validity, but made assessing the validity of the criteria being used for judgements difficult. As an alternative, judges typed up their immediate thoughts about each judgement into a comments box provided after they had made a decision. The volume of notes provided the basis for the opportunistic analysis described here.

Essays and participants

Details of the sample of geography essays and the participants are contained in the report of the main study (Whitehouse & Pollitt, 2012). Suffice it to say, a representative random sample of 564 cleaned essays was judged by 23 examiners and teachers of geography using ACJ.

The current study focused on responses to a compulsory essay question from a summer 2011 AS-level question paper in geography:

> *'Soft engineering is a better river flood management strategy than hard engineering.'*
>
> *Discuss this view.*

The maximum mark for the question was 15; the total maximum for the question paper was 120 marks. There was a generic mark scheme applicable to all of the essays plus a topic-specific mark scheme containing indicative content for each essay. Both mark schemes employed levels of response and are shown in Appendix A. However, the judges in this study were not given the mark scheme to work with.

Making notes on comparative judgements

TAG Development supplied the web-based application used for delivering portable document files (pdfs) of pairs of essays to the judges and calculating the evolving rank order of the quality of the essays. This application and the procedure used are described elsewhere (Kimbell *et al.*, 2009; Pollitt, 2012; Whitehouse & Pollitt, 2012). During the 15 day judging session the judges made 3,519 judgements between pairs of essays or 153 judgements per judge, on average.

On being presented with a pair of essays judges were instructed to make a holistic judgement as to which was the better quality essay. Two importance statements (see Appendix B) and a question guided the judges in making their judgements. The first importance statement was a modified version of the learning aims of a GCE in geography taken from the subject criteria. This provided a link between the rigour of the GCE specification and holistic judgements without the need for a detailed mark scheme. The second importance statement came from the programme of study at key stage 3 for geography in the National Curriculum; there is no equivalent statement for key stage 4 or GCE. The question positioned after the importance statements, was: *Based on these statements, which of the essays shows more evidence of a higher level of development of what is deemed important in Geography?*

After recording their decision as to which was the better essay, the judges were asked to write what was uppermost in their minds immediately after making the judgement. An emphasis was laid on the note being short so as to maintain the momentum of the process of holistic judging.

> ***After** making your decision between two essays, leave a comment about why you made the judgement you made. This comment can be one word, one phrase or one sentence - but if a comment doesn't immediately come to mind, don't worry, leave it and move on to your next judgement.* Extract from *Guidance on making judgements*

Analysis of judges' comments

Valid notes referring to both essays in a judgement were duplicated to give a total of 3,453 notes about the 564 essays. The notes were ordered according to the ranking of the essays to which they applied and then split into three approximately equal sized groups of higher, medium and lower essay qualities; see Table 1[1]. The quality parameters were estimated during the ACJ process and were used to set the rank order of the essays. The upper and lower quality parameters for each group of essays are shown in Table 1.

Table 1: Three groups of notes based on essay quality

Essay quality	No. of notes	No. of essays	Notes per essay	Quality parameter range	
				lower	upper
Higher	1,152	158	7.29	+3.50	+10.80
Medium	1,148	195	5.89	-2.04	+3.49
Lower	1,153	211	5.46	-13.08	-2.03

During the initial readings, the notes were coded for characteristics that included: which essay in the pair the note was about; whether the note was an absolute or comparative statement; and whether the note adopted a positive or negative tone. Subsequently, thematic analysis, a qualitative methodology, was used to detect patterns in the judges' notes. This analytic method was chosen because it is viewed as being flexible, not constrained to any particular theoretical

[1] There are many possible ways in which the essays could have been categorised, however, using a different method would not change the overall findings or conclusions in this report.

framework, relatively easy to use and applicable to a variety of different areas of research. (Braun & Clarke, 2006)

Findings

The findings of this investigation are divided into five sections. The first and second sections describe the characteristics of the notes made by the judges. The emphasis in the second section is on the general tone adopted in the notes. The third section develops the theme of geographical knowledge being linked to progression from lower to higher essay quality. The fourth section considers the terminology used by the judges to describe the skills they observed in the essays and how this terminology may be linked to criteria external to the investigation. The last section presents the findings on how the judges handled superficial features.

Making judgements and making notes

The judges provided valid notes on 92.4% of the 3,519 comparative judgements recorded: Table 2 provides a breakdown of the types of notes made by the judges. One judge was responsible for about a third (68 out of 210 or 32.4%) of the comparative judgements that did not elicit a note and this represented just over half (55.3%) of the comparisons this judge made.

Table 2: Numbers of notes and judgements

Description of note		**Judgements**	
		Number	**Percentage**
Invalid	No note	210	6.0
	System errors	19	0.5
	About decision only	38	1.1
	Sub-total of invalid notes	*267*	*7.6*
Valid	About both essays	201	5.7
	About one essay	3,051	86.7
	Sub-total of valid notes	*3,252*	*92.4*
	Total	**3,519**	

The nineteen invalid notes that recorded system errors related to the uploading of scans of the essays to judges' screens. This problem occurred at the start of the judging session. It was rectified quickly and did not detrimentally affect the judging session. There were another 38 invalid notes that did not refer to either essay. Most of these notes recorded that the judge found making the decision to be difficult. Phrases such as "Very close", "really difficult", "only just", "Too difficult to decide!" and "similar quality" occurred in these notes. One judge noted using a coin toss to decide on the occasions when the judgement was difficult. This was suggested as a strategy in the guidance supplied to the judges so that they did not become enmeshed in one particular decision. The rationale for this guidance was that an essay's final position in the rank order was dependent on a composite of decisions based on professional judgement. One decision in which chance played a role would be more than balanced by the decisions resulting from comparisons with other essays.

The judges were able to make valid notes without any comment on the difficulty of decision-making for 82.7% of all judgements (2,910 out of 3,519). However, the valid notes in which they commented on difficult decisions (342) contained further information about why one essay was chosen over the other. For example:

> *"Both detailed approaches, Essay A slightly edges."* higher quality essay

In a very small number of cases (10) the judge noted that the decision was difficult because the two essays responded to the question in very different ways. For example:

> *“Very difficult to decide. Two different approaches with B linking analysis more directly to place.”* lower quality essay

For most of the judgements with a valid note (3,051 out of 3,252) the note was about one essay.

General tone of judges’ notes

For the remainder of this report percentages refer to notes, not comparative judgements. The total number of notes is 3,453; those 201 notes referring to both essays in a judgement were duplicated (3,252 + 201 = 3,453). Most (89.3%) of the notes the judges made were about the essay that they decided was the better of a pair of essays. More than three-quarters (76.7%) of the notes were not only about the better essay, but also described it in positive terms. About a tenth (10.7%) of the notes was about the losing essay of a pair. Most of these notes were concerned with essays of lower quality (6.9% in the lower quality band and 2.7% in the medium quality band). Table 3 provides a small number of quotes to give a flavour of the tone adopted by the judges. Where the better essay was concerned, the notes were evenly split between making an absolute statement about the essay (35.0% positive and absolute) and comparing it to the poorer essay of the pair (34.3% positive and comparative).

Table 3: Examples of judges’ notes showing coding for tone

Note	Essay quality	Note refers to higher quality essay	Tone
“Has two detailed case studies - at end does examine benefits and issues of each type of engineering”	Higher	yes	absolute positive
“Not much in this one.[2] Discussion perhaps better in winner.”	Medium	yes	comparative positive
“Morer informed - better on physical processes - more scope”	Medium	yes	comparative positive
“clear structure with conclusion, some evidence of understanding of the techniques, pros and cons”	Medium	yes	absolute positive
“broader range of soft and hard engineering techniques”	Lower	yes	comparative positive
“doesn't answer the question”	Lower	no	absolute negative

Making absolute statements when the task was comparative in nature appears counterintuitive. Though the written statement was absolute, there was usually a correspondingly opposite, but

[2] This short sentence was interpreted as meaning that the judge thought the two essays in the comparative judgement were similar, making the decision difficult. This interpretation was supported by the closeness of the two essays in terms of rank order after the ACJ analysis was completed.

unwritten, absolute statement. For example, “includes case studies” may have implied that the other essay in the comparison did not include case studies. Or, a second example: “B has good discussion with relevant examples” may have implied that the other essay in the comparison had a poorer discussion with examples that were either not relevant or not used as competently.

Similar findings about the use of positive tone were reported in a study of the annotations made on scripts from GCSE business studies and GCSE mathematics (Crisp & Johnson, 2007). This may reflect the general philosophy around marking; that it should be “positive rather than negative” and should “credit what candidates know, understand and can do” (Ofqual, 2011, pg.19). This statement is found at the beginning of most mark schemes, including the schemes for AS-level geography that were not supplied to the judges, but is not found in the importance statements that were supplied to the judges.

Geographical knowledge and progression

Geographical knowledge featured greatly in the judges’ notes with 2,946 mentions (some notes contained more than one observation of geographical knowledge). There were four categories of geographical knowledge addressed by the judges. In descending order of importance to the judges, these categories were: case studies and examples; knowledge relating directly to the question as asked; further knowledge relating to the topic; and, wider geographical knowledge.

Knowledge of case studies and examples was noted most frequently by the judges. Whilst case studies and examples have been grouped together, they are different. An example provides an instance of one of the engineering strategies in the question. A case study goes further and “makes links to p lace” as some judges expressed it, giving a location to the example of the strategy and describing the impact of the strategy on the locale. However, for some judges case studies and examples appeared to be interchangeable.

Notes about knowledge, directly related to the question as asked, comprised mentions of the soft and hard engineering strategies or methods. Further topic-related knowledge included notes that specifically mentioned the advantages and disadvantages of the two engineering strategies, geographical processes and the more nebulous words “ideas” and “concepts”. The category of wider geographical knowledge comprised specific items of knowledge including “scale”, “social, economic and environmental aspects” and “sustainability”.

Progression through the essay quality bands was observed through the frequency of notes on the categories of knowledge and the modifying language used by the judges; see Table 4. Judges were most likely to note knowledge that related directly to the question in the lower quality essays. The modifying language noted the lack of knowledge or the presence of knowledge but in a weak state.

> *“Both lack case studies, both limited in their arguments, A slightly edges in their knowledge of defenses.”* a pair of lower quality essays

Medium quality essays had an increased likelihood of judges noting the presence of case studies and examples and further topic-related knowledge. There was a commensurate decrease in the noting of knowledge that was directly related to the question. The modifying language indicated “more” case studies and examples, “accurate” knowledge and “better balance”. Judges tended to note wider geographical knowledge to similar extents in the lower quality and medium quality essays. However, some of the lower quality essays mentioned coastal engineering. Whilst not of direct relevance to the question, knowledge of engineering strategies could still be demonstrated by addressing coastal flooding.

Essays in the higher quality band were more likely to draw notes commenting on the presence of further topic-related knowledge and wider geographical knowledge. For these essays the

Table 4: Likelihood of finding different types of knowledge in each essay quality band

Categories of knowledge	**Essay quality**		
	Lower	**Medium**	**Higher**
Case studies and examples	✓	✔	✔
Directly related to question	✔	✓	✓
Wider geographical	✓	✓	✔
Further topic-related	✓	✔	✔
	Modifying language used		
	attempts, basic, confused, inaccurate, lacks, limited, only, simplistic, unbalanced, vague	(improving) accuracy, balance, better, effective, more, well	clear, confident, detail, excellent, good, higher, quality, range, sophisticated

greatest differentiator was the word "detail". Whilst the frequency of noting case studies was similar to that for the medium quality essays, the modifying language used for the higher quality essays changed to include "range" and "variety" and the use of higher numbers of case studies. They also elicited descriptions such as "sophisticated use of case studies", "High quality of argument" and "winner continues a confident answer throughout".

> "*good knowledge - evaluates 4 detailed case studies.*" higher quality essay

Much of the language used in the notes is also used in the mark schemes (both generic and topic specific, which are presented in Appendix A), despite these mark schemes not being given to the judges. Thus, the lower quality essays were characterised by a relatively high frequency of notes about basic knowledge that was directly related to the question as it was asked. This matches the descriptor in level 1 of the topic-specific mark scheme: "Identifies soft and/or hard engineering strategies"; "Refers to simple reasons why soft engineering is better", and a "Coastal flooding response". The mark scheme also deals with further topic-related knowledge and wider geographical knowledge. Further topic-related knowledge, which included the advantages and disadvantages of soft and hard engineering, is similar to that rewarded in level 2 of the mark scheme. The content of the category of wider geographical knowledge does not find an exact equivalent in level 3 of the mark scheme. However, the phrase "Economies, cultures and environments at scales local and global" and the word "sustainability" appear in the first importance statement; such words were used by the judges to describe higher quality essays.

According to the notes, the presence (or not) of case studies was a key discriminator between qualities of essays. The mark scheme mirrors this. At level 1 responses are not expected to cite case studies, whilst at level 2 "case study material may be used in a descriptive way." At level 3 the generic mark scheme expects "highly detailed accounts of a range of case studies". This chimes with the judges' increased frequency of noting case studies in the higher quality essays and with the modifying language used, such as "detail" and "range". There is a conflict between the generic and topic-specific mark schemes regarding case studies in level 2. The generic mark scheme requires the "detailed use" of case studies, but the topic-specific mark scheme expects only descriptions of case studies. This discrepancy may provide an explanation for the similar frequencies of notes about case studies in the medium and higher quality essays.

Much of the modifying language used in the notes to indicate progression from lower to higher quality essays is also found in the mark scheme. The modifying language used for the lower quality essays, for example, “basic”, “lacks”, and “simplistic” mirrors what is present in the mark scheme at level 1. The importance statements do not indicate how progression is made through the levels of knowledge, or whether one category of knowledge is thought to be more demanding than another. Neither do they directly address the need for a knowledge of case studies.

Observing the demonstration of skills in the essays

There were 2,011 instances in the notes of descriptions of skills demonstrated in the essays. Almost half (43.1%) of these occurred in the higher quality band. Table 5 shows the top five skills noted by the judges for each essay quality arranged in decreasing importance. Four skills recurred for all three essay qualities: understanding, discussing, making an argument and drawing a conclusion. The ranking of these skills was the same for the lower quality essays as for the medium quality essays. Judges tended to describe the skills in the lower quality essays as “basic”, “weak” and partially present with the word “some”. Not drawing a conclusion or not reaching a decision also tended to be observed in lower quality essays, whilst medium quality essays demonstrated the drawing of a conclusion which was either weak or unsupported by evidence. Modifying language used in notes about skills observed in the medium quality essays overlapped with that used for the lower and higher quality essays making it difficult to identify language that was specific to this band.

The observation of the skill of comparing and contrasting in the lower quality essays and the skill of using case studies in the medium quality essays differentiated these two quality bands. How case studies were used in the essays was also the major discriminator between the medium quality essays and the higher quality essays. In the higher quality essays judges tended to note that case studies were used well to support both arguments and conclusions: “case studies integrated into argument”, “well structured argument with case studies” and “case studies develop answer better”.

Table 5: Top five skills noted by judges in each essay quality band

	Essay quality		
Skill	**Lower**	**Medium**	**Higher**
Understanding	1	1	2
Discussing	2	2	3
Making an argument	3	3	4
(not) Drawing a conclusion	4	4	5
Comparing and contrasting	5	--	--
Using case studies	--	5	1
	Modifying language used		
	basic, some, weak	----	clear, depth, focused, supported

Note: ‘using case studies’ is a shorthand form of ‘using case studies and examples’.

The skill of using case studies is emphasised in both the generic and topic-specific mark schemes in levels 2 and 3. Level 3 in particular expects that “case studies are used to make points”. Edwards and Adams (2002) in a comparability study of AS-level geography also found that judges identified the “use of case study material” as a useful script feature on which to base their comparisons.

The skills of understanding, discussing, making an argument and (not) drawing a conclusion were noted for all three qualities of essay. Again, modifying language indicated progression from lower to higher performance. However, only understanding and discussing are directly linked to the assessment objectives and the mark scheme.

Assessment Objective 1: Demonstrate knowledge and understanding of the content, concepts and processes.

Assessment Objective 2: Analyse, interpret and evaluate geographical information, issues, viewpoints and apply understanding in unfamiliar contexts.

According to Bloom's revised taxonomy[3] (Krathwohl, 2002) the skill of comparing and contrasting can be categorised as analysing. Likewise, the skills of discussing, making an argument and drawing a conclusion (deciding or recommending) can be categorised as evaluating. So whilst there were very few notes in which the judges used words derived directly from analysing, applying or evaluating, these skills were still recognised using the terminology shown in Table 5. It is possible to speculate that the latter skills may find their source in the materials associated with teacher standardisation and teacher support.

None of the notes contained the word "interpreting" or related words. Neither did the word "synthesis" appear in any of the notes. This is despite "synthesis" being in the highest level of the generic mark scheme, along with "evaluation" and "assessment". The importance statements do not discuss skills much beyond the use of the words "understanding" and "interpret". There is no attempt to describe progression through the development of skills.

Superficial features of the essays

Superficial features of essays, such as response length and illegibility, did not appear to influence the judges' decisions as they were rarely noted. Any notes about the length of a response (number of sentences, number of paragraphs, or containing the words "long", "short", "brief", or "concise") or expressing a judge's inability to read a student's handwriting were categorised as superficial features.

There were 36 notes that dealt with the length of the response and these comments were raised in three contexts. The first was at the lower end of the lower quality band where the essays were short enough for the words to be counted or the number of sentences to be in single digits and there was no content relevant to the topic. The second was when some judges appeared to conflate length of response with quality of content: "length hence content" (for a lower quality essay) and "length" (for a higher quality essay). However, other judges' comments on the same essays indicated they observed other features that were far less, if at all, related to the length of the response. For example, "uses very vague example. Wider understanding of the issue" and "better attempt at balance in the answer" (for the lower quality essay) and "greater level of detail shown, more precise and discursive" (for the higher quality essay). The third context involved essays that were, perhaps unexpectedly, brief, but that contained evidence of skills. For example, "Brief but better attempt to answer question-brings in elements of location and landuse into the discussion" for a medium quality essay.

There were 16 notes that recorded problems with reading handwriting. For those essays that one or more judges found illegible, there were other judges who did not experience the same difficulty. One judge noted of a medium quality essay: "ERROR diff to read". Other judges reading the same essay noted: "range of supporting evidence", "V difficult to read but is full of information and more support for comparison between 2 types of defence Has a conclusion"

[3] It is reasonable to use Bloom's revised taxonomy here as most of the current assessment objectives are based on its categorisations of learning objectives (Pollitt, Ahmed, & Crisp, 2007).

and “More depth to arguments and has conclusion”. In an operational environment this would be an advantage, with work that one judge finds illegible able to take part in other paired comparisons considered by judges who are able to read the work.

The rarity of notes about legibility is surprising as it does feature in the generic mark scheme in all three marking levels. Legibility is linked with spelling, grammar and punctuation and is considered to contribute to clarity of meaning in the essay. Spelling, grammar and punctuation along with style, flow, fluency clarity and ease of reading all featured more frequently in the notes than did legibility.

Discussion

The current investigation tested the validity of the criteria used in adaptive comparative judgements by thematically analysing the notes made by judges during an experimental study of ACJ. On the whole, judges used the language of the mark scheme and the assessment objectives when describing the knowledge and skills demonstrated in the essays. The majority of notes concerned the better essay in a pair and were couched in positive terms, reflecting the current marking ethos. It was possible to identify patterns in the notes of the types of knowledge and skills that the judges paid attention to and to link these to the generic and topic-specific levels mark schemes (Appendix A). There was also evidence of progression from lower to higher quality essays in the modifying language used in the notes that was mirrored in the mark schemes. Even though the judges were not provided with these mark schemes as part of this study they were still using them implicitly. The language used in the importance statements that the judges were actually provided with was rarely used in their notes.

The judges in this study were examiners or teachers of geography. These roles require them to be familiar with the content of mark schemes, specifications and curricula for their subject. Evidence for this familiarity was present in the language the judges used in the notes about their judgements. It is possible to speculate that in their roles as examiners or teachers the judges had created their own shared construct before taking part in the ACJ experiment. This shared construct uses much of the vocabulary and content from the mark schemes and assessment objectives. However, it omits some criteria, such as legibility and synthesis, and adds in others, such as “making an argument” and “(not) drawing a conclusion”.

A shared construct suggests that the judges in the current study were able to make their decisions in paired comparisons because an established community of practice existed. This community was based around examiner training and published documents such as the mark schemes and specification, which contains the assessment objectives.

A question arises: were the judges using the language of the mark scheme because that is what some of them are trained in and others are familiar with or is the mark scheme an articulation of a set of unchanging criteria that are fundamental to an assessment? If the latter, then these criteria should be innate to a community of practice around the subject. If the former, then the community of practice does not extend beyond the mark scheme itself. Research into teachers’ and students’ views on the stretch and challenge of A-levels and the backwash effect of assessments, suggests a mixture of the two (Baird, Chamberlain, Daly, & Meadows, 2009; Baird, Daly, Tremain, & Meadows, 2009).

If ACJ were to be implemented as an alternative to marking, there would be no established (and used) mark scheme. So the question is raised about how criteria for making holistic judgements using paired comparisons could be shared. The current study suggests that moving an existing assessment from marking to ACJ would not be difficult as shared criteria already exist for judges to use. However, a more cautious approach would have to be adopted with novel assessments that might require that shared criteria for judgement are established alongside the new assessment. Adopting an ACJ system means, in the extreme, there is no need for examiner or teacher training as judgement relies on professional expertise. This would be to ignore the fact that expertise is developed over time and through observation and the sharing of practice (Brooks, 2009; Price, 2005). Shared criteria are best developed by communities of practice if all members of the community are to use them. Under an ACJ style of assessment communities of practice would need to be established and, just as importantly, maintained. This would ensure that there was a pool of judges with shared knowledge available for assessing students' work.

Limitations of the current study

There are a number of limitations to this study, not least the lack of generalisability. This analysis considered only responses to one essay question from one topic within one subject, geography, at one level of qualification, AS. Despite the generic mark schemes for the essay questions within this paper, it is possible the findings may not even be generalisable to other AS-level geography essays.

Note-making is subject to the same sort of flaws as the think aloud method is. In the context of using think aloud to shed light on strategies used during grading, Greatorex and Suto (2008) noted " . . . apparent variation among individuals in the validity of thinking aloud may limit its usage as a 'stand-alone' research method." Note-making as a methodology had three drawbacks. Firstly, despite the simplicity of the instructions, judges interpreted them differently leading to varying amounts of detail and, hence, time devoted to this task. Secondly, it was impossible to know whether a group of judges using similar language had identified exactly the same features within an essay. Yet the cognitive processes the judges used in making judgements were inferred to be the same. Lastly, though there was no direct observation of their decision-making, the judges may still have felt the need to justify their decisions in the notes.

This study identified some of the sources of the criteria on which judges based their decisions. However, it has not demonstrated unequivocally that the ranking of the essays by ACJ was as valid as or any more valid than the ranking achieved by traditional marking. A designed study using judges with different levels of marking experience to rank order scripts by marking and comparative judgement would be needed to find this out. The study should also include an element in which the different rank orders produced are compared by different stakeholders to establish validity.

Conclusions and implications

The results from this study suggest that the teachers and examiners of geography who took on the role of judges were already part of a community of practice that uses shared criteria, giving their decision-making Bramley's psychological validity (2007). This community of practice is based around existing training and information. With an existing assessment, moving from marking to ACJ would be relatively easy as shared criteria exist. However, over time these shared criteria and the community of practice may fade if action is not taken to sustain them. For a new assessment that uses ACJ, the communication of the shared criteria and the establishment or identification of a community of practice would be fundamental to the reliability and validity of that assessment.

References

Adams, R., & Pinot de Moira, A. (2000). *A comparability study in GCSE French including parts of the Scottish Standard grade examination. A study based on the summer 1999 examination. Review of question paper demand, cross-moderation study and statistical analysis of results.*: Organised by Welsh Joint Education Committee and Assessment and Qualifications Alliance on behalf of the Joint Forum for the GCSE and GCE.

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to scaling. *Applied Psychological Measurement, 28*(3), 665-680.

Baird, J.-A., Daly, A., Tremain, K., & Meadows, M. (2009). *Stretch and Challenge in A-Level Examinations: Teachers' Views of the New Assessments.* RPA_09_AD_RP_009 Manchester, UK: AQA Centre for Education Research and Policy.

Baird, J.-A., Chamberlain, S., Daly, A., & Meadows, M. (2009). *Engaging students via backwash: The A-Level stretch and challenge policy.* RPA_09_SC_RP_071 Guildford, UK: AQA Centre for Education Research and Policy.

Bramley, T. (2007), Paired Comparison Methods. In P. E. Newton, J.-A. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (246-294). London: Qualifications and Curriculum Authority.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101.

Brooks, V. (2009). Marking as judgement. *Research Papers in Education, 27*(1), 63-80.

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

Crisp, V., & Johnson, M. (2007). The use of annotations in examination marking: opening a window into markers' minds,. *British Educational Research Journal, 33*(6), 943-961.

D'Arcy, J. (1997). *Comparability Studies between Modular and Non-Modular syllabuses in GCE Advanced Biology, English Literature and Mathematics in the 1996 summer examinations.* Standing Committee on Research on behalf of the Joint Forum for the GCSE and GCE.

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.*

Ericsson, K. A., & Simon, H. A. (1998). How to Study Thinking in Everyday Life: Contrasting Think-Aloud Protocols With Descriptions and Explanations of Thinking. *Mind, Culture and Activity, 5*(3), 178-186.

Fearnley, A. (1999). *A Comparability Study in GCSE Mathematics: syllabus review and cross moderation exercise. A study based on the Summer 1998 examination.* RPA_00_AJF_RC_025 Manchester, UK: AQA (NEAB) on behalf of the Joint Forum for the GCSE and GCE.

Fowles, D. F. (2000). *A Review of the Methodologies of Recent Comparability Studies. Report on an Inter-board Staff Seminar*. RPA_00_DEF_RC_042 Manchester, UK: Northern Examinations and Assessment Board.

Greatorex, J. (2002). Making accounting examiners' tacit knowledge more explicit: developing grade descriptors for an Accounting A-level. *Research Papers in Education, 17*(2), 211-216.

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

Greatorex, J., & Suto, W. M. I. (2008). What do GCSE examiners think of ‘thinking aloud’? Findings from an exploratory study. *Educational Research, 50*(4), 319-331.

Johnson, M., & Nádas, R. (2012). A review of the uses of the Kelly's Repertory Grid method in educational assessment and comparability research studies. *Educational Research and Evaluation: An International Journal on Theory and Practice, 18*(5), 425-440.

Jones, B. E. (Ed.). (1997). *A review and evaluation of the methods used in the 1996 GCSE and GCE comparability studies*. Manchester, UK: Standing Committee on Research on behalf of the Joint Forum for the GCSE and GCE.

Kimbell, R., Wheeler, t., Stables, K., Shepard, T., Martin, F., Davies, D., *et al.* (2009). *e-scape portfolio assessment: phase 3 report* London: Technology Education Research Unit Goldsmiths College University of London.

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy. *Theory into Practice, 41*(4), 213-218.

Newhouse, P. (2011), Comparative Pairs Marking Supports Authentic Assessment of Practical Performance Within Constuctivist Learning Environments. In R. F. Cavanagh & R. F. Waugh (Eds.), *Applications of Rasch Measurement in Learning Environments Research* (141-180). Rotterdam: Sense Publishers.

Nisbett, R. E., & DeCamp Wilson, T. (1977). Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review, 84*(3), 231-259.

Ofqual. (2011) *GCSE, GCE, Principal Learning and Project Code of Practice*. Coventry: Ofqual.

Pashler, H. (1994). Dual-Task Interference in Simple Tasks: Data and Theory. *Psychological Bulletin, 116*(2), 220-244.

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281-300.

Pollitt, A., Ahmed, A., & Crisp, V. (2007), The demands of examination syllabuses and question papers. In P. E. Newton, J.-A. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (166-206). London: Qualifications and Curriculum Authority1.

Price, M. (2005). Assessment standards: the role of communities of practice and the scholarship of assessment. *Assessment & Evaluation in Higher Education, 30*(3), 215-230.

QCA. (2006). *QCA's review of standards*. QCA/06/2374 London: Qualifications and Curriculum Authority.

Suto, I., & Novaković, N. (2012). An exploration of the examination script features that most influence expert judgements in three methods of evaluating script quality. *Assessment in Education: Principles, Policy & Practice, 19*(3), 301-320.

Thurstone, L. L. (1927a). A law of comparative judgement. *Psychological Review, 34*(4), 273-286.

Thurstone, L. L. (1927b). Psychophysical analysis. *American Journal of Psychology, 38*(3), 368-389.

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). The Think Aloud Method. A practical guide to modelling cognitive processes: Academic Press, London.

Whitehouse, C., & Pollitt, A. (2012). *Using Adaptive Comparative Judgement to Obtain a Highly Reliable Rank Order in Summative Assessment*. CERP_12_CW_RP_035 Guildford, UK: AQA Centre for Education Research and Policy.

Appendix A: Mark schemes

Marking for style

Levels Marking – General Criteria

Everyone involved in the levels of marking process (examiners, teachers, students) should understand the criteria for moving from one level to the next – the "triggers". The following general criteria are designed to assist all involved in determining into which band the quality of response should be placed. It is anticipated that candidates' performances under the various elements will be broadly inter-related. Further development of these principles will be discussed during Standardisation meetings. In broad terms the levels will operate as follows:

Level 1: attempts the question to some extent (basic)

An answer at this level is likely to:

- display a basic understanding of the topic
- make one or two points without support of appropriate exemplification or application of principle
- demonstrate a simplistic style of writing perhaps lacking close relation to the terms of the question and unlikely to communicate complexity of subject matter
- lack organisation, relevance and specialist vocabulary
- demonstrate deficiencies in legibility, spelling, grammar and punctuation which detract from the clarity of meaning.

Level 2: answers the question (well/clearly)

An answer at this level is likely to:

- display a clear understanding of the topic
- make one or two points with support of appropriate exemplification and/or application of principle
- give a number of characteristics, reasons, attitudes ("more than one") where the question requires it
- provide detailed use of case studies
- give responses to more than one command e.g. "describe and explain.."
- demonstrate a style of writing which matches the requirements of the questions and acknowledges the potential complexity of the subject matter
- demonstrate relevance and coherence with appropriate use of specialist vocabulary
- demonstrate legibility of text, and qualities of spelling, grammar and punctuation which do not detract from the clarity of meaning.

Level 3: answers the question very well (detailed)

An answer at this level is likely to:

- display a detailed understanding of the topic
- make several points with support of appropriate exemplification and/or appropriate principle
- give a wide range of characteristics, reasons, attitudes, etc.
- provide highly detailed accounts of a range of case studies
- respond well to more than one command
- demonstrate evaluation, assessment and synthesis throughout

- demonstrate a sophisticated style of writing incorporating measured and qualified explanation and comment as required by the question and reflecting awareness of the complexity of the subject matter and incompleteness/tentativeness of explanation
- demonstrate a clear sense of purpose so that the responses are seen to closely relate to the requirements of the question with confident use of specialist vocabulary
- demonstrate legibility of text, and qualities of spelling, grammar and punctuation which contribute to complete clarity of meaning.

Extract from CMI+ annotations

Additional annotations for physical geography

Describes	Landform
Explains	Process
Discusses	Soft engineering
Comment	Hard engineering
To what extent	Sustainability
Cause	Difficulties
Development	Desertification

Extract from Other mechanics of marking

Additional annotations that can be used throughout the question paper

repeated material, vague, not answering question, seen

Marking for question-specific content

1 (c) AO1 – 7, AO2 – 8 (15 marks)

There is a need to make clear why soft engineering strategies are preferred to hard engineering or vice versa. This is the likely route so there should be reference to the advantages of soft engineering and possibly also the disadvantages of hard engineering. There will probably be some description of the relevant strategies that may be adopted.

Alternatively, candidates may disagree with the statement and provide advantages of hard engineering and disadvantages of soft engineering. The final option is to perceive the complementary nature of the two approaches and discuss this aspect.

Advantages of soft engineering are likely to refer to its greater sustainability, its limited interference with a natural system, the ability to improve the environment at times and to work with natural systems so that wetlands and habitats may be restored/created, the relative affordability.

Disadvantages of hard engineering relate to the extent to which there is change to the natural system and questions over its sustainability – the large scale of building dams and their environmental impact, as well as economic and social costs. Similarly, channelization means that the flood risk may be increased downstream and habitats destroyed. **Advantages of hard**

engineering may relate to their effectiveness, especially in the short term, associated schemes for HEP, irrigation which give other advantages.

Disadvantages of soft engineering relate to ineffectiveness in already built-up areas, the fact that flood warnings allow preparation but are not preventing damage from flooding. They will be seen as reducing the scale of risk rather than preventing flooding.

The actual content will depend on the specific strategies considered and whether there is exclusive discussion of soft engineering strategies only. There may be reference to case studies – such as River Quaggy, London, Lincolnshire, Oxfordshire (Cherwell), Ouse, Jubilee River Channel, Carlisle, Three Gorges Dam, Colorado etc.

Level 1 (Basic) 1-6 marks

Identifies soft and/or hard engineering strategies.

Refers to simple reasons why soft engineering is better.

Some use of appropriate terminology present at the higher end.

Coastal flooding response – if relevant, generic aspects.

CMI annotation

- ***L1 Identifies strategies***
- ***L1 Simple reasons given***

Level 2 (Clear) 7-12 marks

Describes strategies and advantages and / disadvantages of soft and / or hard engineering.

Begins to discuss why soft engineering strategies are better (or an alternative option).

Uses strategies to illustrate points – will illustrate one aspect only or with imbalance e.g. advantages of soft engineering may be discussed with no reference to hard engineering.

Case study material may be included in a descriptive way.

Appropriate geographical terminology is used.

CMI annotation

- ***L2 Begins to discuss***

Level 3 (Detailed) 13-15 marks

Clear, purposeful discussion that seeks to put a case for/against soft engineering or is aware of the complementary nature of the strategies.

Advantages and disadvantages of soft and hard engineering are discussed.

Strategies are effectively used to illustrate concepts.

Case studies are used to make points.

Specific terminology is used throughout.

CMI annotation

- ***L3 Purposeful discussion – puts a case***

Appendix B: Importance statements provided to judges

I. Importance Statement for Geography

A student who is developing into a geographer through their course of learning is able to demonstrate that they have

- developed and can apply their understanding of geographical concepts and processes to understand and interpret our changing world
- developed their awareness of the complexity of interactions within and between societies, economies, cultures and environments at scales from local to global
- developed as global citizens who recognise the challenges of sustainability and the implications for their own and others' lives
- improved as critical and reflective learners aware of the importance of attitudes and values, including their own
- become adept in the use and application of skills and new technologies through their geographical studies both in and outside the classroom
- been and are inspired by the world around them, and gain enjoyment and satisfaction from their geographical studies and understand their relevance

Modified from Ofqual's *GCE AS and A level subject criteria for geography*, September 2006

II. The importance of geography

The study of geography stimulates an interest in and a sense of wonder about places. It helps young people to make sense of a complex and dynamically changing world. It explains where places are, how places and landscapes are formed, how people and their environment interact, and how a diverse range of economies, societies and environments are interconnected. It builds on pupils' own experiences to investigate places at all scales, from the personal to the global.

Geographical enquiry encourages questioning, investigation and critical thinking about issues affecting the world and people's lives, now and in the future. Fieldwork is an essential element of this. Pupils learn to think spatially and use maps, visual images and new technologies, including geographical information systems (GIS), to obtain, present and analyse information. Geography inspires pupils to become global citizens by exploring their own place in the world, their values and their responsibilities to other people, to the environment and to the sustainability of the planet.

Geography. Programme of study for key stage 3 and attainment target. (An extract from The National Curriculum 2007) Qualifications and Curriculum Authority, 2007

Based on these statements, which of the essays shows more evidence of a higher level of development of what is deemed important in Geography?