CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Multiple marking methods as alternatives to single marker analytical essay marking: Exploring pairwise comparative judgement, rank ordering and levels-only

Research Report

Emma Walland
Tom Benton
24 March 2026

Author contact details:

Emma Walland & Tom Benton
Research Division
Shaftesbury Road
Cambridge
CB2 8EA
UK

emma.walland@cambridge.org
tom.benton@cambridge.org
https://www.cambridge.org/

As a department of the university, Cambridge University Press & Assessment is respected and trusted worldwide, managing three world-class examination boards, and maintaining the highest standards in educational assessment and learning. We are a not-for-profit organisation.

Cambridge University Press & Assessment is committed to making our documents accessible in accordance with the WCAG 2.1 Standard. We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, contact our team: Research Division

If you need this document in a different format contact us telling us your name, email address and requirements and we will respond within 15 working days.

How to cite this publication:

Walland, E., & Benton, T. (2026). *Multiple marking methods as alternatives to single marker analytical essay marking: Exploring pairwise comparative judgement, rank ordering and levels-only*. Cambridge University Press & Assessment.

Abstract

Single marker analytical marking is a common method used by awarding organisations in England to mark essays written during external examinations. A criticism of this method is the potential lack of reliability. Due to the greater subjectivity of essay marking compared with other forms of assessments, students could get different results if a different marker saw their work. This in turn threatens validity.

In this article, we explore the reliability, predictive value and practicality (in terms of time taken to mark) of single marker analytical marking compared with three alternatives that combine judgments from multiple markers: Rank Ordering, Pairwise Comparative Judgement, and levels-only marking. We compare the methods on English Language essays taken by secondary school students in England as part of their high-stakes examinations. Our aim is to provide assessment practitioners and researchers with preliminary insights on the potential effectiveness of the methods, and to inform further exploration.

Keywords: comparative judgement; mark schemes; multiple marking; reliability of marking.

Contents

Introduction

Essays written during examinations are important assessment tools, allowing us to capture constructs that other types of assessments cannot (Holmes, Black, & Morin, 2017). Essays are used in assessments around the world, and they are noted to have lower marking reliabilities due to the more subjective nature of the marking of extended writing in comparison to shorter answer questions (Holmes et al., 2017; Wheadon, de Moira, & Christodoulou, 2020).

A common method of marking essays, and one which is used in England to mark essays as part of high-stakes examinations taken by secondary school students (GCSEs and A levels)[1], is analytical marking. This requires markers to allocate marks, nested within levels of performance, for different areas of achievement or features of the essay (Meadows & Billington, 2005).

Figure 1 shows an example extract from an analytical mark scheme, used to mark assessment objective (AO) number six for GCSE English Language essays. A similar set of six levels (with mark ranges within each) is used to additionally assess assessment objective number five[2].

[1] GCSEs and A levels are high-stakes exams usually taken in England by 16-year-old and 18-year-old students respectively.

[2] AO5: Communicate clearly, effectively and imaginatively, selecting and adapting **tone, style and register** for different forms, purposes and audiences. Organise information and ideas, using structural and grammatical features to support **coherence and cohesion** of texts.

AO6 - Use a range of **vocabulary and sentence structures** for clarity, purpose and effect, with accurate **spelling and punctuation**.	
Level 4 (13-16 marks)	• An ambitious range of sentence structures is used to shape meaning and create impact. Accurate punctuation is used to enhance clarity and achieve particular effects. • Vocabulary is precise and subtle, expressing complex ideas with clarity. Spelling of irregular and ambitious words is accurate, with very occasional lapses.
Level 3 (9-12 marks)	• A wide range of sentence structures is used for deliberate purpose and effect. Punctuation is consistently accurate and is used to achieve clarity. • Vocabulary is sometimes ambitious and used convincingly for purpose and effect. Spelling, including complex regular words, is accurate; there may be occasional errors with irregular and ambitious words.
Level 2 (5-8 marks)	• A range of sentence structures is used, mostly securely, and sometimes for purpose and effect. Punctuation is generally accurate with occasional errors. • Vocabulary is appropriate and shows some evidence of being selected for deliberate effect. Spelling is generally accurate with occasional errors with common and more complex words.
Level 1 (1-4 marks)	• Simple sentences are used with some attempt to use more complex structures. Some punctuation is used but there is a lack of control and consistency. • Vocabulary is straightforward and relevant with mostly accurate spelling of simple words.
No credit	No response or no response worthy of credit.

Figure 1. An example extract from an analytical mark scheme, used to mark assessment objective number six for GCSE English Language essays.

Marking using complex analytical mark schemes is time and resource consuming in an awarding organisation context. Therefore, only one marker typically marks each essay. This could render the results less reliable, since they are more susceptible to influence by markers' individual preferences, leniencies, or severities, particularly for more subjective subjects such as English. Improved reliability could positively influence the validity of assessments as measures of students' achievement and indications of their potential for future progression. This is because reliability is a necessary, although not sufficient, condition for validity (Kane, 2013).

Efforts to increase marking reliability of essays at awarding organisations in England, for example, have focused on providing in-depth analytical mark schemes with detailed level descriptors. Efforts are made to ensure all markers apply the mark schemes equally through training and marker monitoring. For example, monitoring via seeding is often used. This involves placing exam scripts that have been previously marked by senior markers into more junior markers' allocations so their marking can be continuously checked against known standards. However, efforts to ensure reliable essay marking have not been as successful as hoped and the relationship between the maximum marks per essay and the disagreement among markers remains (Holmes et al., 2017). Making mark schemes more detailed could also negatively influence teaching and learning, narrowing the focus on what will gain marks (Brooks, 2004; Holmes et al., 2017; Wheadon, Barmby, Christodoulou, & Henderson, 2020; Wheadon, de Moira, et al., 2020).

An alternative approach to improving reliability, and which this research explored, is to make use of double or multiple marking. In multiple marking, multiple markers independently assess the same essay, and the final result is a combination of the marks. Although recent studies have found the benefit of double or multiple marking to be quite small in comparison with historical studies, it may have greater impact in certain subjects or with certain items (Benton, 2019; Holmes, Black, & Morin, 2014; Meadows & Billington, 2005). Indeed, studies have found that double marking can improve reliability for items that require an element of subjectivity (Tisi, Whitehouse, Maughan, & Burdett, 2013), such as essays.

For multiple marking to be practical and cost-effective (especially in an awarding organisation context), alternative essay marking methods that consume less time are needed, for example, using less detailed (or more holistic) mark schemes. The time saved could facilitate double or multiple marking at the same financial cost and time as single marker analytical marking.

Holistic marking methods are less detailed and tend to be less time consuming (Meadows & Billington, 2005); however, they are informed by a different paradigm. They involve marking a piece of work based on an overall evaluation, rather than viewing features of the text as separate entities. They are based on the view that the quality of a written text cannot be reduced to a sum of countable features and "can be recognized only by carefully selected and trained readers" (p.79). Research has found that several holistic (or impression) markers could be superior in reliability to one analytic marker (Meadows & Billington, 2005). Therefore, marking methods which are less detailed and more holistic could be combined with multiple marking to lead to improved reliability.

PCJ and RO

Pairwise comparative judgement (PCJ) and rank ordering (RO) have received attention in research for their potential applications in educational assessment (Holmes et al., 2017; Jones & Davies, 2024; Wheadon, Barmby, et al., 2020). PCJ originated from the research of Thurstone (1927) which argued that people are better at making comparisons than absolute judgements. PCJ and RO have since been explored in various educational assessment settings, including as alternatives to marking (Wheadon, Barmby, et al., 2020) and for their potential role in standard maintaining or to assess changes in standards over time for GCSEs and A levels in England (Bramley, Bell, & Pollitt, 1998; Cambridge University Press & Assessment, 2022; Curcin, Howard, Sully, & Black, 2019a). Over time, studies have investigated the reliability, aspects of validity and, increasingly, the efficiency of these methods, however more studies are needed in different contexts (van Daal, Lesterhuis, De Maeyer, & Bouwer, 2022).

In PCJ, markers are presented with many different pairs of essays, and they must choose which one of each pair is better. Each essay must be included in several such comparisons. In RO, markers are presented with several groups (or packs) of essays, from three upwards, and they sort them in order from best to worst. Each essay must be included in several such packs. RO evolved as a potentially more efficient way of doing comparative judgement (Bramley et al., 1998). The data that is generated by PCJ and RO can be converted into a mark for each student.

PCJ and RO differ from analytical marking since they are intended to rely on an overall or holistic impression of the students' essays (rather than an analytical mark scheme). As they make use of multiple judges making multiple judgements, they may capture the varied ways a relevant community of experts understands a construct (Jones & Inglis, 2015). Comparative judgement methods also differ along another dimension, in that they are based on comparisons with other essays rather than marking each essay individually. As noted, in PCJ and RO, markers rank work comparatively, either in pairs of work (PCJ) or in groups of work (RO). This relative nature of marking is argued to be more reliable than absolute judgements (see, for example, Pollitt, 2012). An example extract from the instructions for a PCJ task (used as part of this research) is given in Figure 2, and it illustrates the intended holistic nature of the judgements.

In this approach, you will be presented with **pairs of essays** on the marking software and your task is to select which one demonstrates better performance. What constitutes better performance should be guided by the constructs being assessed (as described by the assessment objectives).

Guidance for making judgements:

- Your judgements should be holistic and intuitive. **Do not re-mark** the essays to come to a decision. Read each essay, think about which one is better and make your decision.
- **Gut reaction/instinct is fine** – you do not need to provide any explanation or justification for your decisions. The fact that, in your opinion, essay A is better than essay B is enough.
- …

Figure 2. An example extract from the instructions for a PCJ task (used in this research).

While PCJ and RO have been shown to yield acceptable reliability as well as evidence towards validity for assessing the intended constructs in various educational assessment settings (Bisson, Gilmore, Inglis, & Jones, 2016; Heldsinger & Humphry, 2010; Holmes et al., 2017; Jones & Inglis, 2015; Verhavert, Bouwer, Donche, & De Maeyer, 2019), questions remain surrounding the clarity of rationales for using comparative judgement in educational assessment (Kelly, Richardson, & Isaacs, 2022). Furthermore, in the context of an awarding organisation, they may not be practical in that the gain in speed of judging may not be sufficient to compensate for the number of times each essay must be judged (Bramley et al., 1998; Bramley & Oates, 2011; Steedle & Ferrara, 2016). Although the methods might require less training and marker monitoring, and more marking could be completed in one session if the methods prove less cognitively demanding (Steedle & Ferrara, 2016), more research is needed to explore the efficiency of the methods in different contexts with different assessment tasks. In awarding organisation specifically, more research is needed

into how much time it takes for PCJ and RO to produce results of comparable or improved reliability and validity in comparison with existing and other alternative marking methods.

A new multiple marking method – levels-only marking (LO)

As part of this research, we developed a new multiple marking method, which we trail alongside PCJ and RO. The new essay marking method, which we called “levels-only” marking (LO), was developed and piloted through a focus group discussion with four senior GCSE English Language markers. Our new LO method could be viewed as sitting somewhere in between a traditional analytical mark scheme and a purely holistic marking method. It is a more holistic and streamlined version of the single marker analytical marking method, whilst still showing elements of analytical marking.

In the new LO method, markers allocate a score for each level for each Assessment Objective (AO), but without determining the precise location of each student within each level. Figure 3 illustrates an example of an extract from the LO mark scheme, for assessment objective six. The main difference from the analytical mark scheme is that the LO mark scheme has no mark ranges within each level. Therefore, the task of marking students’ essays is more straightforward and less detailed. A similar set of six levels (without any mark ranges in each) are also used to mark assessment objective five. For each AO, markers need only choose a level but not the position of the essay within the level (as is done in the existing analytical mark scheme).

AO6 - Use a range of **vocabulary and sentence structures** for clarity, purpose and effect, with accurate **spelling and punctuation**.	
Level 4	• An ambitious range of sentence structures is used to shape meaning and create impact. Accurate punctuation is used to enhance clarity and achieve particular effects. • Vocabulary is precise and subtle, expressing complex ideas with clarity. Spelling of irregular and ambitious words is accurate, with very occasional lapses.
Level 3	• A wide range of sentence structures is used for deliberate purpose and effect. Punctuation is consistently accurate and is used to achieve clarity. • Vocabulary is sometimes ambitious and used convincingly for purpose and effect. Spelling, including complex regular words, is accurate; there may be occasional errors with irregular and ambitious words.
Level 2	• A range of sentence structures is used, mostly securely, and sometimes for purpose and effect. Punctuation is generally accurate with occasional errors. • Vocabulary is appropriate and shows some evidence of being selected for deliberate effect. Spelling is generally accurate with occasional errors with common and more complex words.
Level 1	• Simple sentences are used with some attempt to use more complex structures. Some punctuation is used but there is a lack of control and consistency. • Vocabulary is straightforward and relevant with mostly accurate spelling of simple words.
No credit	No response or no response worthy of credit.

Figure 3. An example of an extract from the LO mark scheme, for assessment objective six.

As noted, we anticipated that LO marking might strike a good balance between analytical marking and more holistic marking methods. Given that the method provides more information about the marking than does PCJ and RO in their current forms, it retains a greater focus on marking transparency. A lack of transparency is a potential concern regarding PCJ and RO as noted by Steedle and Ferrara (2016); van Daal et al. (2022). Given that LO is a streamlined marking method, we expected that the resulting time savings could make multiple marking practical. Our focus group discussion and previous research found that the process of deciding a particular mark within each level (done as part of the analytical mark scheme) is difficult and time consuming for markers (Ahmed & Pollitt, 2011; Fowles, 2009; Hughes & Shaw, 2016; Pinot de Moira, 2011). Thus, removing the need for that could lead to significant time savings which could be reinvested in the multiple marking of each essay. Encouragingly, research by Macinska and Benton (2020) concluded that the information loss by removing the need to allocate marks within levels was marginal for various GCSE and A level subjects.

The mechanism behind the success of PCJ and RO methods, in terms of reliability and aspects of validity, could be that they incorporate elements of both multiple marking as well as holistic marking (Benton & Gallacher, 2018). These elements may also contribute to the success of the proposed LO method. Evidence from the exploratory phase of our research was encouraging as it suggested the new LO method could be positively received by markers as well as save time, making multiple marking feasible. This could lead to associated improvements in the reliability, and therefore validity, of essay results.

Aim

This research aimed to explore the extent to which the multiple marking methods chosen for this study (LO, PCJ, and RO) could produce good quality essay marks, whilst also being practical in terms of the time taken to mark. Whilst the reliability, and aspects of validity, of PCJ and RO have been demonstrated in previous research in various settings (Heldsinger & Humphry, 2010; Holmes et al., 2017; Verhavert et al., 2019), there has been a relative lack of attention to the practicalities in awarding organisation settings (i.e., how long it takes in practice to produce good reliability and validity) as compared with other methods.

Similarly, while streamlining analytical marking schemes into an LO approach should result in marginal information loss, the effectiveness and practicality of this approach, particularly when implemented with multiple marking, is not known.
Thus, in this study, we tested the three multiple marking methods (LO, PCJ and RO) and compared the results with an established method of single marker analytical marking. This allowed us to explore the methods in terms of each of our key variables: reliability, predictive value and practicality. As part of our research, we also explored examiner views and experiences of the different methods, as reported in (Walland, 2022), to factor in qualitative aspects that need considering alongside numerical measures of effectiveness.

Definition of variables

Reliability

Reliability refers to the consistency of a measure and was defined in our study as the extent to which the methods would produce similar essay scores again in the same conditions across different (sets of) markers. Reliability is assessed using different approaches for the different marking methods in our study, due to data availability. Although the methods used to measure reliability share a similar conceptual foundation, this limitation should be borne in mind when comparing the reliability of the methods.
For PCJ and RO, reliability was calculated as the proportion of the variance in essay scores that can be attributed to genuine variances in the quality of essays rather than measurement error (analogous to internal consistency). For LO, reliability was calculated using generalisability theory (Johnson & Johnson, 2009). Specifically, a mixed effects linear model was used to analyse how much of the variance in the total score (the sum of each level for each AO) allocated to each essay was attributable to the essay, the marker and the residual (marker) variance (Benton, 2006).

A reliability of .7 was used as a cut off for assessing what constitutes good reliability in our study. This was influenced by values found in the literature. For the reliability of singe marker analytical marking, there is a wide range of values reported in the literature, depending on the specific assessment and the measure of reliability used (Fowles, 2009; Meadows & Billington, 2005; Tisi et al., 2013). Some studies looked at marker agreement when multiple markers marked the same essay and compared the results with those given by the lead marker. Marker agreement is usually presented using correlation coefficients, and (in this context) values ranging between .50 and .62 might be considered moderate (Holmes et al., 2017). Values higher than this might be considered good, especially in the context of extended writing assessment which tends to produce lower reliabilities and marker agreement. The most comparable approach to how we calculated the reliability of our LO method was found in Benton (2006). The reported reliability of marking ranged from .61 to .73, based on a linear model for an extended writing task and the elements thereof for 11-year-old students. Reports in the literature of reliability for PCJ and RO tend to be very high, with SSRs (scale separation reliability) above .8 and .9 (Holmes et al., 2017). Thus, .7 was considered to be a meaningful representation of good reliability in this context.

Predictive value

We chose predictive value as an additional indicator of the effectiveness of the methods in our context. Predictive value was defined as how well the results derived correlated with other measures of a highly similar construct (as in Benton & Gallacher, 2018). This indicates whether the methods appear to be targeting the same constructs and is, therefore, an indication towards the fitness for the purpose of the methods for assessing the intended constructs. This is sometimes referred to as criterion validity in the literature (e.g., Jones & Inglis, 2015). Other studies have explored aspects of the validity of PCJ in various ways, through, for example comparing it with existing measures (e.g., validated instruments), achievement data (e.g., students' results in a similar course or module), predicted grades or marks resulting from traditional marking rubrics (Bisson et al., 2016; Heldsinger & Humphry, 2010; Jones & Inglis, 2015; Steedle & Ferrara, 2016). Steedle and Ferrara (2016) summarised correlation coefficients used to measure validity reported in previous

comparative judgement studies, and these ranged from non-significant and small, up to over .9. In our study, a cut-off of .7 was used to assess what constitutes good predictive value. According to the literature, .7 is considered a good correlation between predictor and outcome variables in an educational setting (Meadows & Billington, 2005).

In our study, we have deliberately avoided using the terms construct validity, concurrent validity, content validity, criterion validity, predictive validity or convergent validity, although they have similarities with our measure of predictive value. This is because we use predictive value as a relatively simple comparative indicator of the effectiveness of our methods, rather than aiming to comprehensively evaluate the validity of our measures as per validity theory. Predictive value could be considered one small part or indicator towards a validity argument (Kane, 2013; Messick, 1989), and we do not want to overstate what this indicator might show. Furthermore, the aim of our research is to find better methods to assess students' writing. Thus, predictive value was chosen as a more neutral concept, without implying that the marks from single marker analytical marking are the gold standard. This focus on predictive value (rather than validity as a broader concept) should be born in mind when considering the results.

Practicality (in terms of judging or marking time)

For each method, we aimed to explore the effectiveness of the methods when varying degrees of multiple marking was used – i.e., how many markers judged or marked each essay. Since comparative judgment methods have demonstrated evidence of reliability, and aspects of validity, research has increasingly focused on the efficiency of the methods in various settings (Steedle & Ferrara, 2016; van Daal et al., 2022). In our study, we aimed to explore how many judgements per essay would be needed to produce results with high reliability and predictive value for each method, and how much time markers would need. Thus, practicality of the methods was assessed in part in terms of the time taken to produce reliable marks with good predictive value in each marking condition.

A limitation of this measure is that judging time does not entirely capture the practicality of the methods. There are other factors that could affect this which we did not measure as part of our study, such as the time taken for marker training and marker monitoring in each method. As noted by Steedle and Ferrara (2016), whilst PCJ may be more time consuming than other methods, there are potential reductions in terms of training time, marker monitoring time, and not needing to produce a complex marking rubric or mark scheme. We also note that we focus here on practicality of the methods as alternatives to marking in an awarding organisation context. However, the practicality of the methods may vary according to the different settings that the methods are used in, and what their purposes are. For example, Heldsinger and Humphry (2010) argue that comparative judgement may well be more efficient for assessment done by teachers within schools, as the mark schemes (or rubrics) from standardised programs and associated training requirements are complex and impractical for teachers to implement.

Materials and methods

Design

To test the different marking methods against our measures, we used essays written by secondary school students in England, as part of their GCSE examinations (typically taken at age 16). We selected 450 scanned handwritten essays from examinations taken in June 2019 from the examination board Cambridge OCR[3]. The total number of essays comprised three sets of 150 for each marking method (LO, RO and PCJ). The first 150 were chosen randomly (within some parameters[4]) and the other two sets were selected to have the same mark distribution. The number of essays per method was chosen to be in line with previous comparative judgement research (e.g., Bramley & Vitello, 2019), which gives an indication of how many judgements per student are needed to produce reliable results.

We used different essays for each method to ensure accurate judging times. We wished to avoid a situation where markers marked (or made judgements) quicker than they would in practice due to having seen the same essays before as part of trialling other methods. A disadvantage of this approach was that it lowered the power of some of our statistical significance tests, as comparisons between multiple-marking methods were not based on the same sets of essays. Another limitation is the possibility that the characteristics of the essays in each sample influenced the results.

For PCJ, we used a random pairs design[5] with 20 judgements per essay, as per previous research (Curcin, Howard, Sully, & Black, 2019b; Verhavert et al., 2019; Wheadon, de Moira, et al., 2020). For RO, we chose to experiment with each essay in eight packs of ten, as exploratory simulations indicated that increased pack sizes might yield better reliability in less time. Within each method (RO and PCJ), the designs were connected[6]. They were designed to minimise the number of times each participant saw the same essays. No ties were permitted in RO and PCJ.

A potential limitation of the design concerns ecological validity. We cannot be sure of any influence the experimental setting had on marking times and judgements. However, the participants were instructed to mark as they would in a real setting and the data suggested that they completed the tasks conscientiously.

---

[3] The question papers and mark schemes for the June 2019 series can be found on Cambridge OCR's website: https://www.ocr.org.uk/qualifications/gcse/english-language-j351-from-2015/assessment/. Our participants marked essays from the final essay question from the 2019 June series examination paper J351/01 – Communicating information and ideas.

[4] The random sample of essays was selected from students who had also completed the assessment used as the outcome variable for predictive value, for students who used the standard answer booklet (rather than requesting an extra booklet), and for students who had chosen the most popular option from the two choices of essay topic they were given. This made up the large majority of essays.

[5] This means that the pairings were created randomly by the software.

[6] 'Connected' means that every essay is compared to every other one either directly or indirectly. An example of an indirect comparison would be if one pair compares essays A and B, another compares essays B and C, and another compares C and D. A and D are indirectly compared (via B and C).

Participants

Permission was obtained from the examination board Cambridge OCR to recruit markers from the same pool that usually marked the essays. Fifteen participants were recruited following the ethical procedures of the British Educational Research Association [BERA] (2018). They were given detailed information about the study, and the opportunity to ask questions, to make an informed decision about participating. They were able to withdraw up until the data was anonymised and analysed[7].

Participants were selected to be broadly representative of the markers that would usually mark the essays, in terms of their marking experience and performance histories. They had all marked GCSE English Language essays for the most recent three years, and they all had at least four years' English teaching experience. Nine of them had never used any of the new marking methods before. Five of them had some limited experience with comparative approaches (PCJ and RO). The participants were paid for their participation based on a flat rate agreed upfront (not linked to the time they spent on the tasks).

Data collection

The study was carried out in early 2021. Ten participants marked 75 essays each using the LO method, such that each of the 150 essays was marked by five of the ten examiners. All 15 participants ranked ten packs of eight essays each for RO and judged 100 pairs each for PCJ[8].

The participants collectively marked the sets of essays with each method one after another. A counterbalanced design was used such that participants did the methods in different orders to compensate for potential order effects[9].

The participants marked remotely using online, browser-based software. The Cambridge University Press & Assessment CJScaling tool was used for RO and PCJ, and a marking tool created by SR Capture Ltd. was used for LO. The participants found the software easy to use. The environment was similar to usual marking as it was remote marking on the computer, although the software interfaces were simpler than usual. The environment also differed in that it was far less pressured (in terms of both time and responsibility) than marking an actual examination.

The participants were given detailed instructions, marking criteria and software guidance for each method. We gave them this in writing and during an online meeting. For PCJ, they were asked to rate which essay of each pair was better and for RO they were asked to rank packs of ten essays in order from best to worst within each pack. For both PCJ and RO, participants were told not to remark the essays but to judge them using a professionally formed holistic judgement in relation to the AOs (covering the main constructs of content, organisation and technical accuracy). Their judgements were converted into numerical

[7] A full ethical review was not required, as per the policy of the Cambridge University Press & Assessment ethics committee.

[8] After they had marked with each method, we collected their views and experiences through surveys and interviews, and the findings for RO and PCJ are reported in Walland (2022).

[9] There was no evidence that the speed of marking depended upon the order of the methods (see Appendix A).

measures of quality for each essay, to enable comparison with the marks generated by the other methods.

For LO, they were given the simplified LO mark scheme we created which consisted of allocating each essay two scores - a score from zero to six for the first AO (content and organisation) and a score from zero to four for the second (technical accuracy). The scores were added together to create a final score out of ten[10].

The time taken to mark each essay, pair or pack was automatically recorded by the marking software[11]. Participants were informed and consented to being timed prior to participating. They were given a deadline and could complete the tasks at a time of their choosing during the time period. They were explicitly instructed to go at their own usual pace for marking as one of our key aims was to work out accurate time estimates. For comparison purposes, we obtained marks allocated to the essays using single marker analytical marking from the 2019 examination series data. Marking times for analytical marking were estimated from the data from the marking software used in the examination series.

Appendix C shows examples of the initial data generated by the PCJ and RO tasks, including the time taken to judge, for illustrative purposes.

Data cleaning and preparation

All data was cleaned and analysed using SAS Enterprise Guide (version 7.1) and R statistical software (version 3.5.0). Prior to conducting analysis relating to our main variables - reliability, predictive value and judging time - we created measures from the marking results and analysed judge and script fit. These two aspects are discussed in turn.

Creating measures

PCJ and RO led to raw data consisting of: for PCJ, which essay from each pair was chosen as the better essay; and for RO, the position of each essay within each pack of ten in order from best to worst. This data was converted into a numerical measure of quality for each essay. For PCJ and RO, we prepared the data for analysis by converting the comparisons data into measures using Plackett-Luce models in R (Turner, van Etten, Firth, & Kosmidis, 2018). The models created measures that placed each essay on the latent trait scale of perceived quality. This created a set of results to use for further analysis, which can be treated as a form of exam marks. An example of the resulting data converted into measures is shown in Appendix D for illustrative purposes.

For LO marking, the scores given for each essay for each AO were added together to form a mark out of ten (illustrated in an example in Appendix E). The marks for single marker analytical marking were obtained from the data from the 2019 actual examination series and no cleaning or preparation was required.

---

[10] The mark schemes for LO marking are in Appendix B.

[11] This was measured as the time between opening each essay, pair or pack, and submitting the final result, and in some cases, with breaks in between removed when participants paused and restarted the task later.

Next, we created measures for different numbers of judgements per essay for PCJ, RO and LO[12]. This was to enable analysis of how reliability and predictive value changed as a function of the number of judgements per essay and, therefore, the time taken to complete marking. For PCJ and RO, we developed a method to successively remove some of the participants' judgements to create measures ranging from four judgements per essay (for PCJ) and two judgements per essay (for RO) up to the maximum. Originally, our maximum was 20 judgements per essay for PCJ and eight for RO, but this was reduced slightly to 18.67 and 7.5 respectively due to us removing the data from a misfitting judge.

To create these measures, we wrote code in R that successively removed packs in RO and pairs in PCJ one at a time from the data so that the minimum number of packs or pairs that any essay was included in remained as high as possible. As the essays removed at each stage were randomly selected (within the parameters), the entire removal process (i.e., removing packs one by one) was repeated 100 times. The figures we calculated for reliability and predictive value represent the average of the values across all repetitions. An example of this dataset for PCJ is shown in Appendix F for illustrative purposes.

For LO marking, we also developed a method to combine the marks from multiple markers into results for different numbers of judgements per essay. Each essay had been judged by five markers, and this represented the maximum. The scores for this were simply the average across all five participants' marks. Single marking represented the minimum and we created measures of reliability and predictive value for single marking by taking the average of these across 100 different replications of randomly choosing single marks for each essay. For double to quadruple (or four marker multiple) marking we also used the average results across 100 replications of randomly choosing combinations of participants' marks for each essay. For example, the predictive value correlation coefficients we calculated represent the average of the coefficients for those 100 random combinations. An example of this dataset for LO is shown in Appendix G for illustrative purposes.

Judge and script fit

Judge fit and script fit are routinely calculated for PCJ and RO, and we also explored judge fit for our new LO method. Judge fit explores whether the markers were marking in a similar way to each other and to the marks that were allocated to the essays using single marker analytical marking. For RO and PCJ, our analysis of infit and outfit mean square statistics (see, for example, Gill, Bramley, & Black, 2007) found that one marker had worrisome fit statistics as per Linacre (2002), and their data was removed from the dataset. We fortunately had sufficient data for this marker's data to be removed without limiting our results unduly. In a real setting, the marker would have been identified earlier and supported or stopped from making judgements. For LO marking, we looked at judge fit by exploring the correlation between the marks markers awarded each essay and the total mark (sum) across the other four markers that marked the same essay. We also looked at the correlation of each marker's mark with the actual mark allocated to the essay in the real

[12] Measures by the number of judgements were not explored for analytical marking because, although multiple marker analytical marking would likely lead to improvements in reliability, this would not be feasible in practice given the time and resource constraints.

examination series. Based on this, we retained all data for LO as there were no outliers. Appendix H shows the judge fit analyses in more detail.

Script fit explores how well the individual judgements on each essay for PCJ and RO fit the statistical model used to combine the judgements into single measures of essay quality. The fit of essays within the PCJ and RO analysis was explored using the infit and outfit mean square statistics (see Wright & Masters, 1990 for details of the calculation steps). We also explored the impact of removing essays with severe levels of misfit as defined by having either infit or outfit mean square values greater than two (Linacre, 2002). As the impact on our results was very limited, and scores would be needed for all essays in a real-world marking scenario, all essays and decisions were retained. For a detailed analysis of script fit, please see Appendix I.

Data analysis

As noted, our main variables of analysis were reliability, predictive value and practicality (judging or marking time). Part of our analysis included comparing how these changed as a function of the number of markers or judgements per essay so that we could explore the impact of different forms of multiple marking on the key variables. The analyses carried out for each main variable are explained in the following sections.

Reliability

As noted previously, reliability was defined as a measure of the consistency of the marking methods. For PCJ and RO, reliability was calculated using a formula which calculates the proportion of the variance in essay scores (from PCJ and RO) that can be attributed to genuine variances in the quality of essays rather than measurement error (analogous to internal consistency). The formula we used can be found in Appendix J. The reliability of PCJ is known to be influenced by factors such as the number of judgements per pair (see, for example, Verhavert et al., 2019), and research studies tend to use SSR (scale separation reliability) to assess the reliability of PCJ and RO. High reliability values are usually reported (Heldsinger & Humphry, 2010; Holmes et al., 2017). Inter-rater reliability has also been used in the context of PCJ, for example, Jones, Swan, and Pollitt (2015), and was found to be high in the context of assessing mathematics problem solving.

For LO, reliability was calculated using generalisability theory (Johnson & Johnson, 2009). Specifically, a mixed effects linear model was used to analyse how much of the variance in the total score (the sum of each level for each AO) allocated to each essay was attributable to the essay, the marker and the residual (marker) variance (Benton, 2006). The variance attributed to the essay only, as a proportion of overall variance in individual LO scores, was used to indicate the reliability of marking, as shown in Appendix K. Other potential reliability measures such as inter-rater reliability (e.g., kappa or weighted kappa) were not used in this case, as our methods were multiple marking methods making use of a combination of marks across different markers, and different assessment objectives, rather than a single level from a single marker. Our method to measure the reliability of LO also enabled better comparisons with the reliability of PCJ and RO, as the methods are conceptually similar.

Since we used the existing marks that had been given to these essays using analytical marking in the past, we did not have data that could allow us to estimate marking reliability

for single marker analytical marking. Instead, values from the literature for similar assessments were used for a very rough indication, and this limitation should be borne in mind when interpreting the findings. There is a wide range of values for the reliability of marking English essays reported in the literature, depending on the specific assessment and measure used (Fowles, 2009; Holmes et al., 2014; Meadows & Billington, 2005; Tisi et al., 2013). The most comparable approach to how we calculated the reliability of our LO method was found in Benton (2006) although a limitation is that the data was for younger students' writing (at age 11). The reported reliability ranged from .61 to .73, based on a linear model for an extended writing task and the elements thereof.

Predictive value

As noted previously, in our study, predictive value was defined as a measure of how well the results derived from each marking method correlated with other measures of a similar construct. We calculated the predictive value for LO, RO and PCJ at different numbers of judgements per essay. We also calculated the predictive value of single marker analytical marking for comparison. For all methods, the predictive value was determined by calculating Spearman's rank-order correlation coefficients. This is a suitable choice for ordinal or ranked data and matches other similar studies (see e.g., Holmes et al., 2017).

For the outcome variable, we used the students' marks on a highly similar essay question for another of the GCSE English Language examinations, which is taken by students in the same examination period (but in a different exam)[13]. The essay uses the same mark scheme and assesses the same assessment objectives. Before running the predictive value correlations, we examined the dataset for potentially influential outliers, but decided to retain all data points as the impact was insignificant, as shown in Appendix L.

A potential limitation of our measure of predictive value was that the outcome variable comes from an analytical single-marking approach rather than a holistic approach. To address this limitation, further research was conducted to see whether the predictive value of PCJ changed if the outcome variable was also scored using PCJ rather than with single marker analytical marking (Ireland, Walland, & Benton, 2022). We found that there was no difference in the predictive value.

Judging time

Judging time (or the time taken to mark) is important to measure to gather more information to assess how practical the methods might be in a real marking situation. However, as noted in the Aim section, there are other factors that would affect practicality linked to the marking context, such as training time. To calculate judging time for LO, RO and PCJ, we used the robust mean marking time per essay, pack and pair respectively using all markers' data. We

[13] For predictive value, the outcome variable (the second essay) was the essay scores for the last essay question of the J351/02 examination paper – Exploring effects and impact. As students have a choice between two equivalent essay questions for this examination paper, the scores for whichever option they chose were used. Two of the students in our sample omitted the essay question for the second essay (used as the outcome) and were, therefore, excluded from the predictive value calculations. The question papers and mark schemes for the June 2019 series can be found on OCR's website. https://www.ocr.org.uk/qualifications/gcse/english-language-j351-from-2015/assessment/.

chose the robust mean as it weights the data such that less plausible values, perhaps relating to a marker opening an essay and then taking a break, are given less weight[14]. For comparison purposes, an estimate of the robust mean time required to mark a single essay via single marker analytical marking was also calculated, by analysing data from actual marking, as shown in Appendix M.

Thereafter, the robust mean marking times were used to calculate how much judging time would be needed in total for each method, for various numbers of judgements per essay. Based on this, we estimated total judging time for 150 and 1000 essays for illustration, to explore this aspect of the practicality of the methods in the context of awarding organisation marking.

Results and discussion

In this section, we present the results integrated with our discussion of them. We begin with reliability and predictive value, followed by judging time. Lastly, we bring the variables together into an overall comparison of the different methods.

Reliability and predictive value

The reliability and predictive value for various numbers of judgements per essay (indicating the extent of multiple marking) for each method are shown in Table 1[15].

[14] The robust means were calculated with an intercept-only robust regression of the total times for each task for each method, using the *MASS* package in R. The robust mean was preferred to the median as the latter may slightly underestimate marking times by failing to give enough weight to genuinely hard-to-mark essays.
[15] As noted, for single marker analytical marking, estimates from the literature were used for a rough indication of reliability.

Table 1. The reliability and predictive value of each method for various numbers of judgements per essay. Values in bold indicate high reliability and predictive value (above .70). Multiple predictive values for the single marker analytical approach are provided depending upon which sample of essays calculations were completed for.

Method	Judgements per essay	Reliability	Predictive value (r_s)	N
Single marker analytical	1	.61 to **.73** (estimation from literature)	.64 (PCJ sample) .60 (RO sample) .63 (LO sample)	150 (PCJ sample) 149 (RO sample) 149 (LO sample)
PCJ	18.67	**.87**	**.74**	150
	18	**.86**	**.74**	150
	16	**.84**	**.74**	150
	14	**.82**	**.73**	150
	12	**.79**	**.72**	150
	10	**.75**	**.71**	150
	8	.68	.69	150
	6	.58	.66	150
	4	.38	.61	150
RO	7.5	**.92**	.64	149
	7	**.91**	.63	149
	6	**.90**	.62	149
	5	**.88**	.61	149
	4	**.85**	.60	149
	3	**.80**	.58	149
	2	.69	.54	149
LO	5	**.90**	**.73**	149
	4	**.87**	**.73**	149
	3	**.84**	**.71**	149
	2	**.78**	.68	149
	1	.63	.62	149

Note: Judgements per essay for RO means how many packs of ten each essay was in.

Table shows that, within each method (PCJ, RO and LO), the reliability and predictive value increased as the number of judgements per essay increased. This was expected and indicates the potential value of multiple marking for improving reliability and predictive value. In simple terms, when multiple markers mark the essays, we can achieve better results than if we have a single marker for each essay. When there was enough multiple marking (or sufficient numbers of judgements per essay), PCJ, RO and LO showed high reliability (above .70). This high reliability for RO and PCJ matches figures reported elsewhere, which tends to be above .80 and .90 (Bisson et al., 2016; Heldsinger & Humphry, 2010; Holmes et al., 2017). For LO, reliability was very good from double marking (two judgements per essay) upward and reached up to .90 for five judgements per essay. This further illustrates the value of multiple marking for reliability. As noted in the Aim section, each method assessed reliability in a different (although conceptually similar) way due to data availability and the differing natures of the marking methods. This should be borne in mind when considering the reliability of the different methods. Due to the reliabilities being measured on different

scales, the reliability statistics for PCJ and RO are not statistically comparable to those of LO and single marker analytical marking. That is, some of the variations in reliability coefficients across different methods may be due to differences in the nature of the scales from, such as, how much they stretch out the scores given to students at the top and bottom of the scales.

Turning to predictive value, the results for LO and PCJ were good when there were sufficient judgements per essay, but the results for RO were not as good and did not reach above .70 even with the maximum number of judgements per essay. However, this could be partly due to sampling; the sample of essays used for RO had a lower predictive value for single analytical marking than the other two methods to start with (i.e., the correlations between the original marks on the two sets of essays was lower for this group).

The predictive value for the various methods can be statistically compared and the differences were tested for statistical significance. For the differences between the methods and single marker analytical marking, we used an online calculator to compare correlations from dependent samples[16]. To compare the predictive values of PCJ, RO and LO amongst each other (i.e., from independent samples), Fisher's r to Z transformation was used (Sheskin, 1997; Wuensch, 2019; Zar, 1999)[17].

The results urge caution in overinterpreting the differences in predictive value among PCJ, RO and LO, as the differences among them were not statistically significant. However, when comparing with single marker analytical marking, LO triple marking (and upwards) and PCJ with 18.67 judgements per essay were statistically significantly better (PCJ: p =.004, z = 2.70, LO: p = .009, z = -2.36). Other than this, the differences between the methods were not statistically significant. This could be influenced by the fact that different essays were used for LO, RO and PCJ, which greatly reduces the statistical power of such comparisons. This was a known drawback of our design but was deliberately chosen to ensure that the judges involved in each method has not marked the same essays beforehand which might have affected their marking speeds. That is, we prioritised accurate measurement of the speed of each method over power for statistical comparisons of predictive value.

Judging time

One of our aims was to explore the practicalities of the marking methods in an awarding organisation setting, and comparing judging time is one indicator towards this. PCJ is known to achieve good reliability, and some studies have found evidence towards validity (in terms of the correlations of the results with measures of similar constructs) (Bisson et al., 2016; Heldsinger & Humphry, 2010), although it is noted to be time consuming to achieve good results (Coertjens, 2017). Recent studies have turned attention to focusing on the practicality

[16] https://www.psychometrica.de/correlation.html#dependent.

[17] This is appropriate for comparing correlations when the sample size is greater than 10 and Spearman's rho is less than .90. The formula for standard error of the Fisher-Z difference is, however, modified slightly for Spearman correlations, to 1.060/n-3. The alpha was set at .05. For full details, see Appendix N.

of the methods in different settings, which has been identified as a key research question (Bramley & Oates, 2011; van Daal et al., 2022).

Including judging time as a variable in our analyses, along with our analysis of reliability and predictive value, enabled us to gather some initial data towards evaluating this. We compared this with the results we can achieve using LO with multiple marking or single marker analytical marking. As noted previously, however, judging time is not the only consideration for evaluating the practicality of the methods. There are other factors, such as training time and marker monitoring time, which could vary among the different methods and in different settings.

First, we present the robust mean times taken to mark an essay (for LO), a pair of essays (for PCJ) and a pack of ten essays (for RO) in Table 2. Thereafter, we use these times to produce estimated judging times for each method for the different numbers of judgements per essay. We use this, along with predictive value and reliability, to inform an overall evaluation of the methods (in the following section).

Table 2. Time per task for each method.

Method	Robust mean time per task	Standard error	N
PCJ	3.65	0.07 (4s)	1400
RO	30.58	1.66	112
LO	3.13	0.08 (5s)	750
Single marker analytical	8.92	0.90	5431

Note: 'Task' refers to judging a pair of essays for PCJ, sorting a pack of ten essays for RO, and marking one essay for LO and analytical marking. All times are in minutes unless stated otherwise.

Overall evaluation

A key aim of our study was to evaluate the marking methods in an awarding organisation setting by considering reliability, predictive value as well as practicality in terms of the time taken to mark (judging time). Table 3 shows the data sorted in ascending order of judging time. Bold and shading indicate instances of high reliability (above .70), high predictive value (above .70), and/or potentially feasible judging time (similar or less than single marker analytical marking). If we look for methods with bold and shading for all three criteria, this gives us an indication of which multiple marking methods were the most effective for GCSE English Language essay marking according to our data, as alternatives to single marker analytical marking in an awarding organisation context.

Table 3. Overall evaluation of the judging time, reliability and predictive value of the methods.

	Method	Judgements per essay	Judging time for 1000 essays (hours)	Judging time for 150 essays (hours)	Reliability	Predictive value
Similar or less judging time than single analytical marking	**LO**	1	**52**	**8**	.63	.62
	RO	2	**102**	**15**	.69	.54
	LO	2	**104**	**16**	**.78**	.68
	PCJ	4	**122**	**18**	.38	.61
	Analytical	1	**149**	**22**	*.50 to .**73***[18]	.60 to .64
	RO	3	**153**	**23**	**.80**	.58
	LO	3	**157**	**23**	**.84**	**.71**
Much more judging time than single analytical marking	**PCJ**	6	183	27	.58	.66
	RO	4	204	31	**.85**	.60
	LO	4	209	31	**.87**	**.73**
	PCJ	8	243	37	.68	.69
	RO	5	255	38	**.88**	.61
	LO	5	261	39	**.90**	**.73**
	PCJ	10	304	46	**.75**	**.71**
	RO	6	306	46	**.90**	.62
	RO	7	357	54	**.91**	.63
	PCJ	12	365	55	**.79**	**.72**
	RO	7.5	381	57	**.92**	.64
	PCJ	14	426	64	**.82**	**.73**
	PCJ	16	487	73	**.84**	**.74**
	PCJ	18	548	82	**.86**	**.74**
	PCJ	18.67	568	85	**.87**	**.74**

Looking at all the variables together, we can see that in this context and according to our data, LO marking with three markers per essay appears to be the most effective method as a potential alternative to single marker analytical marking, with LO double marking a close second.

For PCJ and RO, when reliability and predictive value are both high, the judging time is much higher than analytical marking and would require greater time and resource to implement, which may not be practical in our context. However, there were differences in the original predictive values of the methods for analytical marking so some of the differences could be due to the samples. As noted, the time taken to mark is also not the only measure of practicality, as there are other factors such as the time required for marker training and marker monitoring that would need to be factored in.

[18] Estimates from the literature.

The data from Table 3 is illustrated in Figure 4 (predictive value) and Figure 5 (reliability). The values written on the lines represent the number of judgements per essay. For predictive value, single marker analytical marking (analytical) is also shown for comparison for each separate set of essays (the brown points). As noted, a limitation is that the reliabilities for LO are not entirely statistically comparable with RO and PCJ as they represent the reliabilities of scores on different scales.

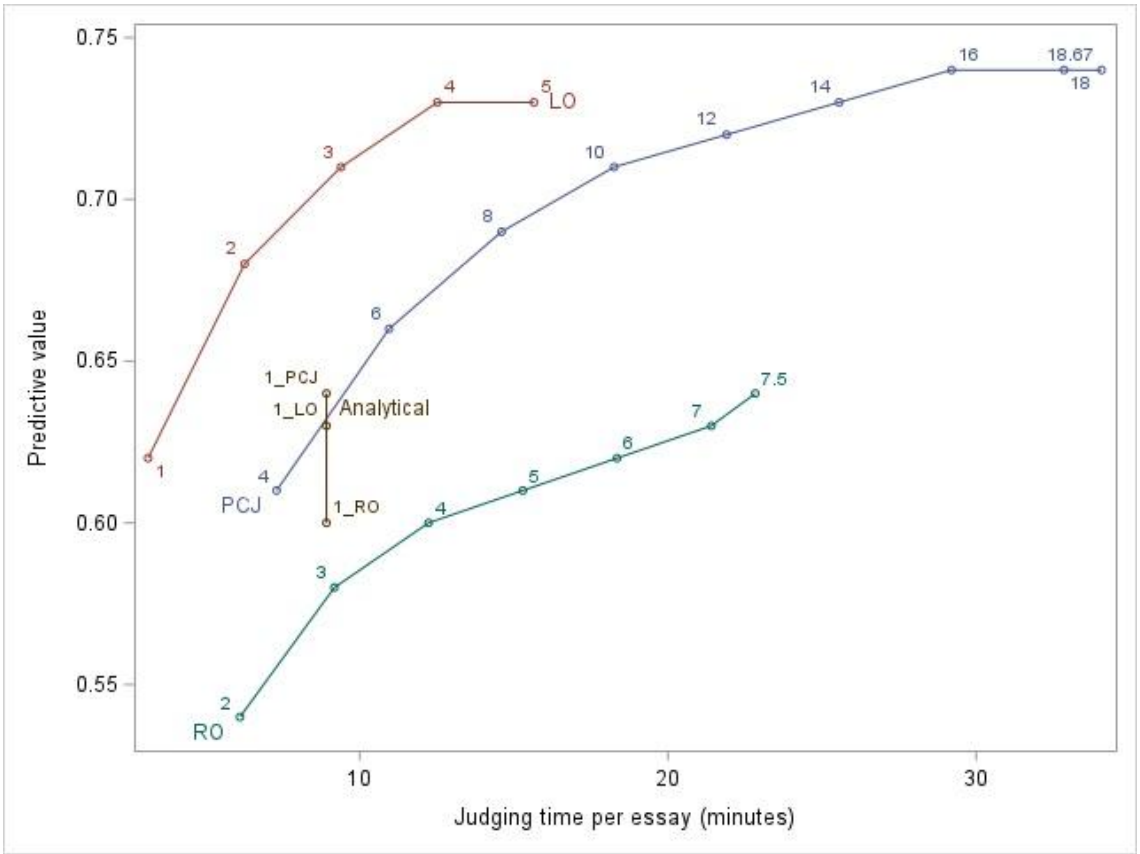

Figure 4. The judging time and predictive value of the different methods for different numbers of judgements per essay.

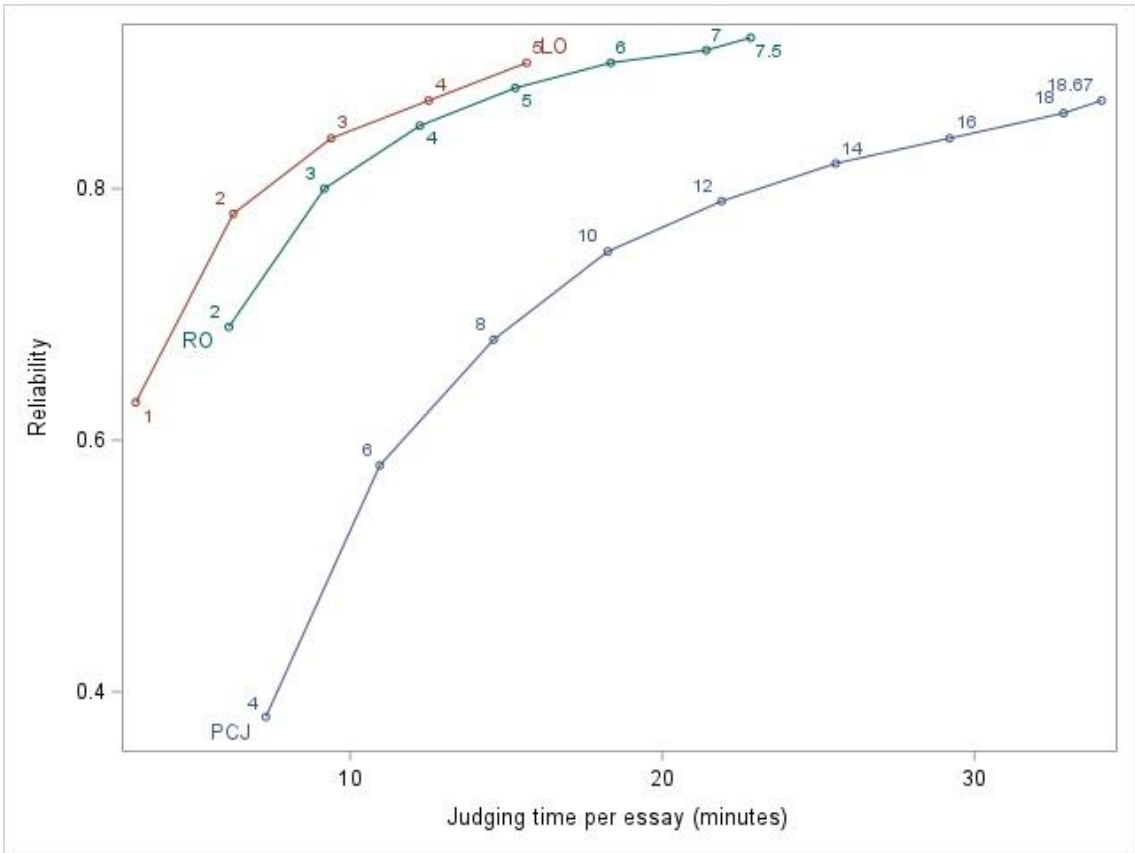

Figure 5. The judging time and reliability of the different methods for different numbers of judgements per essay.

Figure 4 and Figure 5 illustrate how reliability and predictive value changed as we increased the number of judgements per essay and, therefore, the judging time. Looking at the graph allows us to see that in this context, for any fixed amount of human resource (in this case, judging time per essay), the LO approach to marking provided higher levels of both reliability and predictive value, according to our data.

Overall, our data indicates the possibility of LO marking with three markers per essay as a potential alternative to single marker analytical marking in this context, with LO with double marking a close second. Both methods provided high values for reliability and predictive value whilst requiring similar amounts of marking time to existing analytical marking. We recommend further research and exploration of the potential of these methods in this and different contexts, including further measures of reliability, validity and practicality.

Limitations

Our research presents an initial evaluation towards understanding the potential of these methods for essay marking in an awarding organisation setting, limited to the small set of variables considered. It is worth raising other factors that potentially need considering alongside reliability, predictive value and judging time if these methods were to be used in practice. For example, factors like stakeholder reactions to the methods, and the impact on teaching and learning (including considerations of feedback). Some of these factors have been explored further in a related publication (Walland, 2022), which reports the views and

experiences of our participants on using PCJ and RO. For LO, feedback from markers suggests that the LO method was positively perceived, however, there was some concern about the relative lack of detail provided in the final scores produced by the method in contrast with single marker analytical marking. This requires further investigation to understand the role of this detail in the examining process and what functions it fulfils, while acknowledging that assessment design often involves trade-offs.

As noted, we do not see predictive value as synonymous with validity of the methods for assessing essay quality, and further evidence on the validity of the methods for their intended purposes would help evaluate them further.

Another limitation of the current study is that we only have results for a single assessment (secondary school English essays) in the context of external marking in an awarding organisation in England. As such, we cannot be certain that the pattern of results we have seen in this present study would be repeated across a wider range of extended writing assessments or in different contexts. It appears likely that the LO method would be applicable to different assessments that have extended writing items marked with single marker analytical marking, and could be beneficial in terms of reliability and validity without requiring a major increase in required resource. However, it would be worthwhile to explore this empirically in further experiments.

As we have pointed out, our measures of reliability for each method were different, which limits the strength of our comparisons. We did not have measures of the reliability of single marker analytical marking, and so had to make use of similar estimates from previous literature, which were in different contexts. We had different essays marked in each method (although with matched mark distributions), to help ensure accurate marking times, and this could have affected our results and made our statistical comparisons of predictive value weaker. There is the possibility that the characteristics of the essays in each sample influenced the results.

Our measures of judging or marking time are not perfect measures but should be seen as estimates. In our study, marking times (as well as other variables) could have been influenced by the marking being done in a research setting rather than an actual high-stakes marking setting. The marking times for single marker analytical marking were not collected as new data in our study, but were estimated from data from marking software. The judging times are a limited indication of practicality of the methods, as they do not account for other time spent during the process, and they focus on application of the methods in a particular setting (an awarding organisation). Future research is needed to explore additional aspects of practicality, thorough comparison with factors such as training time and marker monitoring time.

Conclusion

Essays written as part of examinations are important assessment tools that allow us to capture constructs that other assessments cannot. However, they are potentially vulnerable to lower marking reliabilities due to increased subjectivity associated with marking extended responses. In this study, we explored different possible multiple marking methods for external marking of secondary school English Language essays in an awarding organisation

context as potential alternatives to single marker analytical marking. A key focus was considering reliability and predictive value together, alongside the practicality of the methods in terms of the time taken to mark. Our data showed that, in this context, three marker multiple marking using LO could potentially be achieved using the same time as single marker analytical marking, whilst yielding significantly better predictive value (of .71) and high reliability (of .84). Beyond the impact on reliability and predictive value, the presence of multiple marking could also reassure students and teachers that their marks are not due to a particular marker's views.

Our study adds to existing knowledge about the potential of PCJ and RO as alternatives to marking. Despite evidence of reliability and aspects of validity, the practicality of PCJ and RO in this context is still a concern, as well as potential issues with the methods running counter to trends towards increased transparency in educational assessment (Steedle & Ferrara, 2016; van Daal et al., 2022). There remains concern about whether PCJ judgements partially capture construct-irrelevant features, such as handwriting, (Chambers, 2022), and questions exist about how using PCJ as a marking method might influence question design (Jones & Inglis, 2015).

We further add to the literature through considering measures of marking quality as well as practicality in terms of time taken to mark for different methods in this context. Additionally, our study contributes a new method - LO - to the literature.

In our study, while PCJ produced results with good reliability and predictive value when there were sufficient numbers of judgements per essay, our data showed that it was more time consuming than single marker analytical marking and LO. According to our analyses, if the same amount of time were used for PCJ as analytical marking, the predictive value would be similar, but the reliability would be poor. For RO, if the same time as single marker analytical marking were used, the predictive value would be similar, and it would have good reliability. Overall, however, LO produced the highest gain in predictive value and reliability using a similar amount of judging time according to our analysis. In the context of awarding organisation external essay marking, the LO method may strike a balance between purely holistic methods and elaborate marking rubrics, and function as a practical way to implement multiple marking (should that be desired). The different methods explored in this study give different information, and whether or not they would be suitable in various settings depends on the context and intended purposes of their application (Landrieu, De Smedt, Van Keer, & De Wever, 2022; van Daal et al., 2022).

Our design has several limitations, which means that these conclusions should be viewed as preliminary. We recommend that further exploration into the LO method in various contexts would be worthwhile for both researchers as well as assessment designers considering ways to improve essay marking reliability or to implement multiple marking in a practical way.

Acknowledgements

We would like to thank all the participants who took part in this research. We would also like to thank those who reviewed our report for their valuable feedback.

References

Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice, 18*(3), 259-278. doi:10.1080/0969594X.2010.546775

Benton, T. (2006). *Exploring the importance of graders in determining pupils' examination results using cross-classified multilevel modelling*. Paper presented at the European Conference on Educational Research, Geneva. https://www.nfer.ac.uk/media/1722/ekq01.pdf

Benton, T. (2019). Which is better: One experienced marker or many inexperienced markers? *Research Matters: A Cambridge Assessment Publication, 28*, 2-10. doi:10.17863/CAM.100393

Benton, T., & Gallacher, T. (2018). Is comparative judgement just a quick form of multiple marking. *Research Matters: A Cambridge Assessment Publication (26)*, 22-28. Retrieved from https://www.cambridgeassessment.org.uk/Images/514987-is-comparative-judgement-just-a-quick-form-of-multiple-marking-.pdf

Bisson, M.-J., Gilmore, C., Inglis, M., & Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Mathematics Education, 2*(2), 141-164. doi:10.1007/s40753-016-0024-3

Bramley, T., Bell, J. F., & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone's paired comparisons. *Education Research and Perspectives, 2*, 1-23.

Bramley, T., & Oates, T. (2011). Rank ordering and paired comparisons - the way Cambridge Assessment is using them in operational and experimental work. *Research Matters: A Cambridge Assessment Publication, 11*, 32-35. doi:10.17863/CAM.100519

Bramley, T., & Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 26*(1), 43-58. doi:10.1080/0969594X.2017.1418734

British Educational Research Association [BERA]. (2018). Ethical Guidelines for Educational Research, fourth edition. Retrieved from https://www.bera.ac.uk/researchers-resources/publications/ethical-guidelines-for-educational-research-2018

Brooks, V. (2004). Double marking revisited. *British Journal of Educational Studies, 52*(1), 29-46. doi:10.1111/j.1467-8527.2004.00253.x

Cambridge University Press & Assessment. (2022). Exam board uses comparative judgement to help set GCSE grade boundaries. Retrieved from https://www.cambridgeassessment.org.uk/news/exam-board-uses-comparative-judgement-to-help-set-gcse-grade-boundaries/

Chambers, L., & Cunningham, E. (2022). Exploring the validity of comparative judgement: Do judges attend to construct-irrelevant features? *Frontiers in Education, 7*.

Coertjens, L., Lesterhuis, M., Verhavert, S., Van Gasse, R., & De Maeyer, S. (2017). Judging texts with rubrics and comparative judgment: Taking into account reliability and time investment. *Pedagogische Studien, 94*(4), 283–303.

Curcin, M., Howard, E., Sully, K., & Black, B. (2019a). *Improving awarding: 2018/2019 pilots*. Retrieved from https://www.gov.uk/government/publications/improving-awarding-20182019-pilots

Curcin, M., Howard, E., Sully, K., & Black, B. (2019b). Improving awarding: 2018/2019 pilots. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/851778/Improving_awarding_-_FINAL196575.pdf

Fowles, D. (2009). How reliable is marking in GCSE English? *English in Education, 43*(1), 50-67. doi:10.1111/j.1754-8845.2009.01032.x

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 Annual Conference, London. https://www.cambridgeassessment.org.uk/Images/109760-an-investigation-of-standard-maintaining-in-gcse-english-using-a-rank-ordering-method.pdf

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher, 37*(2), 1-19. doi:10.1007/BF03216919

Holmes, S., Black, B., & Morin, C. (2014). *Review of Double Marking Research*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/605661/2014-02-14-review-of-double-marking-research.pdf

Holmes, S., Black, B., & Morin, C. (2017). *Marking reliability studies 2017: Rank ordering versus marking – which is more reliable?* Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/859250/Marking_reliability_-_FINAL64494.pdf

Hughes, S., & Shaw, S. (2016). Why do so few candidates score 4 out of 8 on this question? The issue of under-used marks in levels-based mark schemes. *Research Matters: A Cambridge Assessment Publication*(21). doi:10.17863/CAM.100340

Ireland, J., Walland, E., & Benton, T. (2022). *The predictive value of Pairwise Comparative Judgement [forthcoming]*. Internal research report. Cambridge University Press & Assessment.

Johnson, S., & Johnson, R. (2009). Conceptualising and interpreting reliability. Retrieved from https://core.ac.uk/download/pdf/4160331.pdf

Jones, I., & Davies, B. (2024). Comparative judgement in education research. *International Journal of Research & Method in Education, 47*(2), 170-181. doi:10.1080/1743727X.2023.2242273

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educational Studies in Mathematics, 89*(3), 337-355. doi:10.1007/s10649-015-9607-1

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing Mathematical Problem Solving Using Comparative Judgement. *International Journal of Science and Mathematics Education, 13*(1), 151-177. doi:10.1007/s10763-013-9497-6

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. doi: www.jstor.org/stable/23353796

Kelly, K. T., Richardson, M., & Isaacs, T. (2022). Critiquing the rationales for using comparative judgement: a call for clarity. *Assessment in Education: Principles, Policy & Practice, 29*(6), 674-688. doi:10.1080/0969594X.2022.2147901

Landrieu, Y., De Smedt, F., Van Keer, H., & De Wever, B. (2022). Assessing the quality of argumentative texts: Examining the general agreement between different rating procedures and exploring inferences of (dis)agreement cases. *Frontiers in Education, 7*. doi:10.3389/feduc.2022.784261

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2). Retrieved from https://www.rasch.org/rmt/rmt162f.htm

Macinska, S., & Benton, T. (2020). *The usefulness of detailed marks within the levels of levels-based mark schemes*. Retrieved from https://www.cambridgeassessment.org.uk/Images/593879-the-usefulness-of-detailed-marks-within-the-levels-of-levels-based-mark-schemes.pdf

Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability*. Retrieved from https://research.aqa.org.uk/sites/default/files/pdf_upload/CERP_RP_MM_01052005.pdf

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd ed)* (pp. 13-103). New York: Macmillan.

Pinot de Moira, A. (2011). *Levels-based mark schemes and marking bias*. Retrieved from AQA Centre for Education Research and Practice: https://research.aqa.org.uk/sites/default/files/pdf_upload/CERP_RP_APM_24112011.pdf

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281-300. doi:10.1080/0969594X.2012.665354

Sheskin, D. (1997). *Handbook of Parametric and Nonparametric Statistical Procedures*: CRC Press.

Steedle, J. T., & Ferrara, S. (2016). Evaluating Comparative Judgment as an Approach to Essay Scoring. *Applied Measurement in Education, 29*(3), 211-223. doi:10.1080/08957347.2016.1171769

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review, 34*(4), 273.

Tisi, J., Whitehouse, G., Maughan, S., & Burdett, N. (2013). *A review of literature on marking reliability research*. Retrieved from https://www.nfer.ac.uk/media/2002/mark01.pdf

Turner, H., van Etten, J., Firth, D., & Kosmidis, I. (2018). Introduction to PlackettLuce. Retrieved from

https://cran.rstudio.com/web/packages/PlackettLuce/vignettes/Overview.html#:~:text=A%20classic%20model%20for%20such,which%20the%20choice%20is%20made.

van Daal, T., Lesterhuis, M., De Maeyer, S., & Bouwer, R. (2022). Editorial: Validity, reliability and efficiency of comparative judgement to assess student work. *Frontiers in Education, Volume 7*. doi:10.3389/feduc.2022.1100095

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice, 26*(5), 541-562. doi:10.1080/0969594X.2019.1602027

Walland, E. (2022). Judges' views on pairwise Comparative Judgement and Rank Ordering as alternatives to analytical essay marking. *Research Matters: A Cambridge University Press & Assessment publication, 33*, 48-67. doi:10.17863/CAM.100427

Wheadon, C., Barmby, P., Christodoulou, D., & Henderson, B. (2020). A comparative judgement approach to the large-scale assessment of primary writing in England. *Assessment in Education: Principles, Policy & Practice, 27*(1), 46-64. doi:10.1080/0969594X.2019.1700212

Wheadon, C., de Moira, A. P., & Christodoulou, D. (2020). *The classification accuracy and consistency of comparative judgement of writing compared to rubric-based teacher assessment*. Retrieved from https://doi.org/10.31235/osf.io/vzus4

Wright, B., & Masters, G. (1990). Computation of OUTFIT and INFIT Statistics. *Rasch Measurement Transactions, 3*(4), 84-85.

Wuensch, K. L. (2019). Comparing correlation coefficients, slopes, and intercepts. Retrieved from http://core.ecu.edu/psyc/wuenschk/docs30/CompareCorrCoeff.pdf

Zar, J. H. (1999). *Biostatistical analysis*. India: Pearson Education.

Appendices

Appendix A – Order effects

Table 4 shows the median times per task for each judge for each method, in the order in which they completed them. Values in bold are results higher than the group median and those not in bold are lower than the group median. Based on this, we can see that in general, judges who marked slower or quicker than the group median tended to across all of their methods regardless of the order. There was only one judge (Judge 5) that was slower for their first method, but faster than the group median for their remaining two.

Table 4. Median time per task for each participant per method, in the order in which they completed them.

Judge	Method 1	Median time	Method 2	Median time	Method 3	Median time
4	LO	**3.40**	PCJ	**5.58**	RO	**35.23**
3	PCJ	**4.53**	RO	**44.23**	LO	**7.88**
8	PCJ	**5.87**	RO	**40.43**	n/a	n/a
9	PCJ	**6.46**	RO	**94.83**	n/a	n/a
11	RO	**44.29**	PCJ	**4.24**	n/a	n/a
6	LO	1.92	PCJ	2.34	RO	28.36
7	RO	5.75	PCJ	1.40	n/a	n/a
10	RO	25.03	LO	1.68	PCJ	0.64
15	RO	8.83	PCJ	0.98	n/a	n/a
1	PCJ	2.28	LO	2.35	RO	23.38
13	PCJ	n/a	LO	2.13	RO	n/a
14	PCJ	2.43	LO	1.68	RO	28.78
5	RO	**30.16**	PCJ	2.94	LO	2.02
12	LO	**2.97**	RO	28.18	PCJ	**4.26**
2	PCJ	**3.84**	RO	25.98	LO	**4.50**

Figure 6 shows the same data graphically in terms of the relationship between the order in which judges experienced a particular method and how quickly they completed their tasks. This reveals no obvious evidence of the speed at which judges completed their tasks being dependent upon the order in which the undertook them.

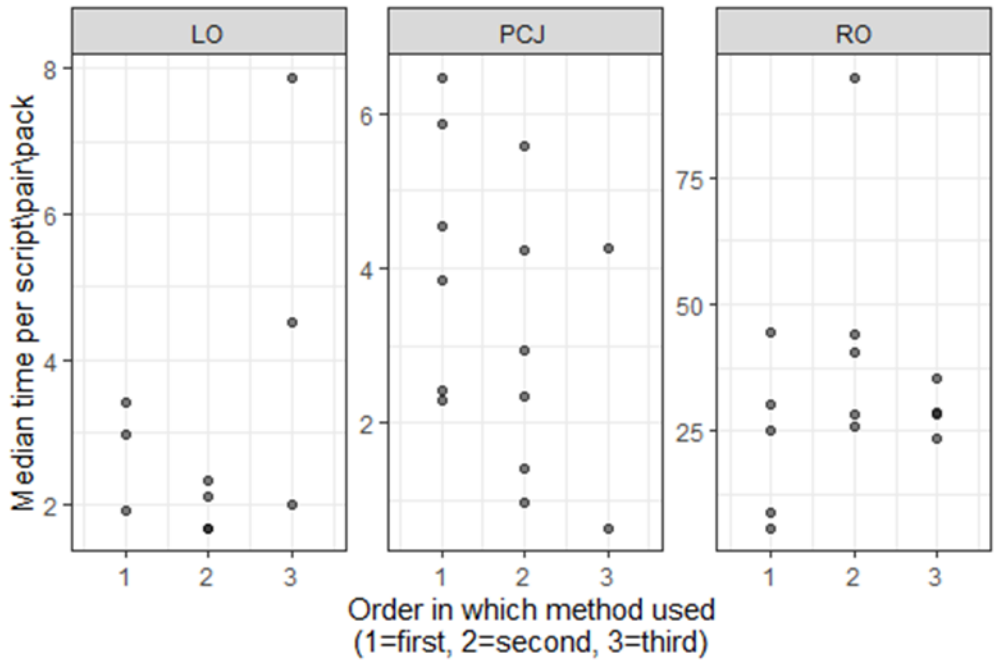

Figure 6. Median marking times per judge for each method by the order in which they trialled the methods.

Appendix B – LO mark scheme

AO5 - Communicate clearly, effectively and imaginatively, selecting and adapting **tone, style and register** for different forms, purposes and audiences. Organise information and ideas, using structural and grammatical features to support **coherence and cohesion** of texts.	
Level 6	• The form is deliberately adapted to position the reader, showing a sophisticated control of purpose and effect. • Tone, style and register are ambitiously selected and deployed to enhance the purpose of the task. • There is a skilfully controlled overall structure, with paragraphs and grammatical features used to support coherence and cohesion and achieve a range of effects.
Level 5	• The form is confidently adapted and shows a secure understanding of purpose and audience. • There is a sustained use of tone, style and register to fulfil the purpose of the task. • There is a controlled overall structure, with paragraphs and grammatical features used to support coherence and cohesion and achieve particular effects.
Level 4	• The form is adapted to show a clear understanding of purpose and audience. • Tone, style and register are chosen to match the task. • There is a well–managed overall structure, with paragraphs and grammatical features used to support coherence and cohesion, and sometimes for effect.
Level 3	• The form is sustained and shows clear awareness of purpose and audience. • Tone, style and register is appropriate for the task, with some inconsistencies. • There is a clear overall structure, with paragraphs and grammatical features used, mostly securely, to support coherence and cohesion.
Level 2	• The form, which is mostly appropriate for purpose and audience, is generally maintained. • There is an attempt to use a tone, style and register appropriate to the task. • There is some evidence of overall structure, with some use of paragraphs and grammatical features to support coherence and cohesion.
Level 1	• There is some attempt to use a form appropriate for purpose and audience. • There is a limited attempt to use a tone, style and register appropriate for the task. • There is some attempt to structure the response, with limited evidence of paragraphs or grammatical features to support coherence and cohesion.
No credit	No response or no response worthy of credit.

AO6 - Use a range of **vocabulary and sentence structures** for clarity, purpose and effect, with accurate **spelling and punctuation**.	
Level 4	• An ambitious range of sentence structures is used to shape meaning and create impact. Accurate punctuation is used to enhance clarity and achieve particular effects. • Vocabulary is precise and subtle, expressing complex ideas with clarity. Spelling of irregular and ambitious words is accurate, with very occasional lapses.
Level 3	• A wide range of sentence structures is used for deliberate purpose and effect. Punctuation is consistently accurate and is used to achieve clarity. • Vocabulary is sometimes ambitious and used convincingly for purpose and effect. Spelling, including complex regular words, is accurate; there may be occasional errors with irregular and ambitious words.
Level 2	• A range of sentence structures is used, mostly securely, and sometimes for purpose and effect. Punctuation is generally accurate with occasional errors. • Vocabulary is appropriate and shows some evidence of being selected for deliberate effect. Spelling is generally accurate with occasional errors with common and more complex words.
Level 1	• Simple sentences are used with some attempt to use more complex structures. Some punctuation is used but there is a lack of control and consistency. • Vocabulary is straightforward and relevant with mostly accurate spelling of simple words.
No credit	No response or no response worthy of credit.

Appendix C - Initial PCJ and RO data examples

Table 5. Example of the initial data derived for PCJ.

Judge	Pair	Essay	Order presented to judge	Rank order	Time per essay	Total task time (per pair) (s)
1	1354	195	A	1	134.5	269
1	1354	201	B	2	134.5	269
1	526	261	A	1	83.5	167
1	526	195	B	2	83.5	167
1	175	280	A	1	118	236
1	175	197	B	2	118	236
1	484	166	A	1	95	190
1	484	203	B	2	95	190
…						

Table 6. Example of the initial data derived for RO.

Judge	Pack	Essay	Order presented to judge	Rank order	Time per essay	Total task time (per pack) (s)
1	66	305	A	7	193	1244
1	66	421	B	8	145	1244
1	66	358	C	4	136	1244
1	66	339	D	10	92	1244
1	66	386	E	3	99	1244
1	66	406	F	6	186	1244
1	66	346	G	9	95	1244
1	66	302	H	2	110	1244
1	66	396	I	5	115	1244
1	66	432	J	1	73	1244
…						

Appendix D - PCJ data example (after measures)

Table 7. Example of the PCJ data after measures were created, and showing the variables used for predictive value calculations. The RO dataset was in the same format.

Essay	Measure	Standard error	Paper 1 essay mark (single analytical marking)	Paper 2 essay mark (single analytical marking)
166	2.009712	0.687325	29	30
195	0.105497	0.569734	25	40
197	-1.05263	0.571604	25	22
201	-0.43098	0.537729	21	24
203	-2.1765	0.668467	21	12
261	1.472102	0.644816	23	37
280	1.202336	0.683443	27	31
…				

Appendix E – Initial LO data example

Table 8. Example of the initial LO dataset.

Marker	Essay	AO5 score	AO6 score	Total score	Total time (s)
1	1	3	3	6	653
1	21	3	2	5	717
1	22	6	3	9	653
1	23	6	4	10	284
1	24	5	4	9	1443
1	25	4	2	6	652
1	26	5	3	8	762
…					

Appendix F - PCJ data example for different numbers of judgements per essay

Table 9. Example of the dataset created to enable analysis of the variables of interest for different numbers of judgements per essay for PCJ.

Repetition	Number of judgements per essay	Essay	Measure	Standard error
1	4	151	-1.88741	1.963405185
1	4	152	0.273439	1.658727041
1	4	153	0.671145	1.502643077
… repeat for each essay				
1	6	151	-2.51354	1.712832014
1	6	152	-0.13176	1.399864003
1	6	153	1.531225	1.213554528
… repeat for each essay				
1	8	151	-3.619	1.712809544
… repeat for each essay, for up to 18 judgements per essay, and up to 100 repetitions.				

Appendix G - LO data example for double to quadruple marking

Table 10. Example of the LO dataset, for double to quadruple marking.

Repetition	Judgements per essay	Essay	LO score
1	2	1	7
1	2	2	5
1	2	3	7
1	2	4	8
… repeat for each essay, for up to 4 judgements per essay, for 100 random combinations.			

Appendix H – Judge fit

The results for judge fit (rounded to two decimal places) are shown in Table 11 for PCJ, Table 12 for RO and Table 13 for LO. The median time per pack, pair and essay are also shown for reference. Infit and outfit mean square values greater than two for PCJ and greater than one for RO are flagged in bold.

Table 11. PCJ judge fit statistics and median time per pair.

Judge	Number of pairs	Infit	Outfit	Median time / pair (minutes)
1	100	0.88	0.59	2.27
2	100	0.76	0.51	3.84
3	100	0.74	0.46	4.53
4	100	0.66	0.44	5.58
5	100	0.92	0.89	2.94
6	100	0.78	0.62	2.34
7	100	1.20	1.60	1.40
8	100	0.59	0.39	5.87
9	100	0.59	0.35	6.46
10	100	0.76	0.54	0.64
11	100	0.92	0.87	4.24
12	100	0.89	0.83	4.26
13	100	1.55	**2.31**	5.49
14	100	0.88	0.70	2.43
15	100	0.65	0.46	0.98

Table 12. RO judge fit statistics and median time per pack.

Judge	Number of packs	Infit	Outfit	Median time / pack (minutes)
1	8	0.96	0.75	23.38
2	8	0.85	0.61	25.98
3	8	0.91	0.69	44.23
4	8	0.76	0.59	35.23
5	8	0.95	0.71	30.16
6	8	0.85	0.68	28.36
7	8	**1.15**	0.97	5.75
8	8	0.64	0.41	40.42
9	8	0.74	0.53	94.83
10	8	0.89	0.77	25.03
11	8	0.93	0.76	44.29
12	8	0.85	0.59	28.18
13	8	**1.34**	**1.52**	30.08
14	8	0.84	0.65	28.77
15	8	**1.09**	1.00	8.83

Table 13. LO judge fit statistics and median time per essay.

Judge	N	Mean total	SD total	r_s with others	r_s with original mark	Median time / essay (minutes)
1	75	6.43	1.48	0.67	0.61	2.35
2	75	6.52	1.69	0.86	0.71	4.50
3	75	5.93	2.33	0.85	0.81	7.88
4	75	6.49	1.89	0.86	0.81	3.40
5	75	6.71	1.67	0.61	0.69	2.02
6	75	6.75	1.09	0.78	0.66	1.92
10	75	6.97	1.44	0.72	0.67	1.68
12	75	5.97	1.80	0.76	0.71	2.97
13	75	6.29	1.97	0.63	0.59	2.13
14	75	6.31	1.57	0.80	0.76	1.68

Appendix I – Script fit

For the PCJ analysis, four essays (out of 150) were identified as severely mis-fitting. In every, case the high level of misfit could be traced to a single pairwise decision deemed to be highly unlikely given the estimated measures for the two essays in the pair. However, removing these essays from the analysis made very little difference to the analysis of predictive value. Specifically, removing the four essays reduced the Spearman correlation between essay measures for essay 1 (the predictor variable) and essay 2 (the outcome variable) from 0.744 to 0.735. Further analysis explored the impact of removing all severely mis-fitting individual decisions (i.e., essay pairs) from the data before estimating essay measures[19]. A total of 12 such decisions were identified (out of 1400). However, removing these decisions from the data had almost no impact on predictive value with the relevant Spearman correlation now being 0.743.

Measures of infit and outfit (mean square) for each essay were also calculated for the RO data. These were based upon comparing the observed rank of each essay within each pack to its expected value and standard deviation given the estimated essay measures. These calculations were facilitated by simulation[20]. Five essays were identified as severely mis-fitting. However, removing these essays from the analysis made very little difference to predictive value. Specifically, removing these essays only changed the Spearman correlation between essay measures from the RO exercise on essay 1 (the predictor variable) and analytical marks for essay 2 (the outcome variable) from 0.639 to 0.649. Removing individual mis-fitting rankings from the data (specifically the 28 most mis-fitting rankings of particular essays within particular packs) was also found to have almost no influence on the estimate of predictive value (revised Spearman correlation of 0.632).

[19] These were identified by the absolute value of their standardised residuals (Z) as defined by Wright and Masters, 1990. Any decisions where the absolute value of this residual exceeded two were defined as severely mis-fitting.

[20] The formulae from Wright and Masters (1990) still apply. However, rather than being calculated mathematically, the values of E and W in their formulae were calculated by simulating 1,000 replications of all the within-pack rankings based on the estimated script measures, and then calculating the mean and variance of the rankings across the simulations.

Wright, B., & Masters, G. (1990). Computation of OUTFIT and INFIT Statistics. *Rasch Measurement Transactions, 3*(4), 84-85. https://www.rasch.org/rmt/rmt34e.htm

Appendix J – Reliability (PCJ and RO)

The reliability for RO and PCJ was calculated with the following formula, where MSE is the mean square of the standard errors associated with each estimate of essay quality and SD is the overall standard deviations of essay quality measures:

$$R = \frac{SD^2}{MSE + SD^2}$$

Some comparative judgement research uses the Scale Separation Reliability (SSR) as the reliability measure. This is the ratio of true variance to observed variance and refers to the proportion of variance that can be attributed to differences between essays. The formula is as follows, where RMSE is the root mean square of the standard errors:

$$SSR = 1 - \left(\frac{RMSE}{SD}\right)^2$$

In our research, we used an alternative form of reliability, after having calculated both and checking to see which was more accurate. The check uses a simulation based on us knowing (i.e., simulating) the true essay measures and then compares the two possible estimates of reliability to the squared correlation between the true and the estimated values. For both the PCJ and RO data, it was found that our reliability formula yielded more accurate results.

For completeness, a comparison of the estimated SSRs against the estimates of reliability we have used in our analysis (labelled “R”) is shown in Table 14. The negative value of the SSR for small numbers of pairs per essay in the PCJ method confirms why the method was not used.

Table 14. Comparison of SSR and reliability according to our method.

Method	Judgements per essay	Reliability
PCJ	18.67	.84 (SSR) **.87 (R)**
	18	.84 (SSR) **.86 (R)**
	16	.81 (SSR) **.84 (R)**
	14	.78 (SSR) **.82 (R)**
	12	.73 (SSR) **.79 (R)**
	10	.66 (SSR) **.75 (R)**
	8	.53 (SSR) **.68 (R)**
	6	.27 (SSR) **.58 (R)**
	4	-.62 (SSR) **.38 (R)**
RO	7.5	.91 (SSR) **.92 (R)**
	7	.90 (SSR) **.91 (R)**
	6	.88 (SSR) **.90 (R)**
	5	.86 (SSR) **.88 (R)**
	4	.82 (SSR) **.85 (R)**
	3	.74 (SSR) **.80 (R)**
	2	.56 (SSR) **.69 (R)**

Appendix K – Reliability (LO)

The mixed effects linear model used to calculate the reliability of LO marking was represented by the following equation:

$$y_{ik} = s_i + m_k + e_{ik}$$

where the essay score (y) for essay 'i' by marker 'k' equals the effect of the essay (s) on the score for essay 'i', plus the effect of the marker leniency or severity (m) for marker 'k', plus the effect of marker variance (e) for essay 'i' for marker 'k'. Using this model, reliability is defined as the proportion of overall variance in individual LO scores attributed to the performance on the essays only. The parameter estimates resulting from the model for single LO marking are shown in Table 15.

Table 15. Covariance parameter estimates for the mixed linear model, for single LO marking.

Covariance Parameter	Estimate	% Variance
Essay (s)	1.96	63.29
Marker (m)	0.10	3.10
Residual (e)	1.04	33.61

To calculate the reliability for different numbers of judgements per script, the following formula was used (for n number of markers, where 'm' refers to the marker and 'e' the residual parameter estimates):

$$Rel_n = \frac{var(s)}{var(s) + \frac{var(m)}{n} + \frac{var(e)}{n}}$$

As n increases (and we average across scores), the variances due to markers should decrease in a predictable way.

Appendix L – Outliers for predictive value

We inspected scatter plots of the outcome versus predictor variables and used Studentized residuals[21] to explore outliers in the PCJ, RO and LO datasets, in preparation for the predictive value analysis. We also inspected the plot for the analytical marks for the combined dataset. The plots are shown in Figures 7 to 10, and the outliers according to Student residuals are encircled.

For PCJ and RO, we found no essays with Student residuals greater than |3|. The scatter plots indicated that there may be some potential outliers. We experimented with removing some of the potential outliers identified, but as this had marginal to no effect on the results, we decided to retain all essays in the predictive value analysis. Furthermore, retaining all essays reflects the fact that, in an operational setting, it would be necessary to provide scores for all essays.

For LO quintuple marking, the scatter plot was inspected, and two essays looked to be potential outliers that had student residuals greater than |3|. Removing them did have a small impact on the correlations for LO and analytical marking, causing them to increase slightly. However, we opted to retain them for consistency and because they appeared to due to the students not completing their essays. For the analytical marks, there was only one outlier identified using Student residuals, but it was retained for consistency.

[21] The residual divided by an estimate of its standard deviation. This is a technique used to detect outliers.

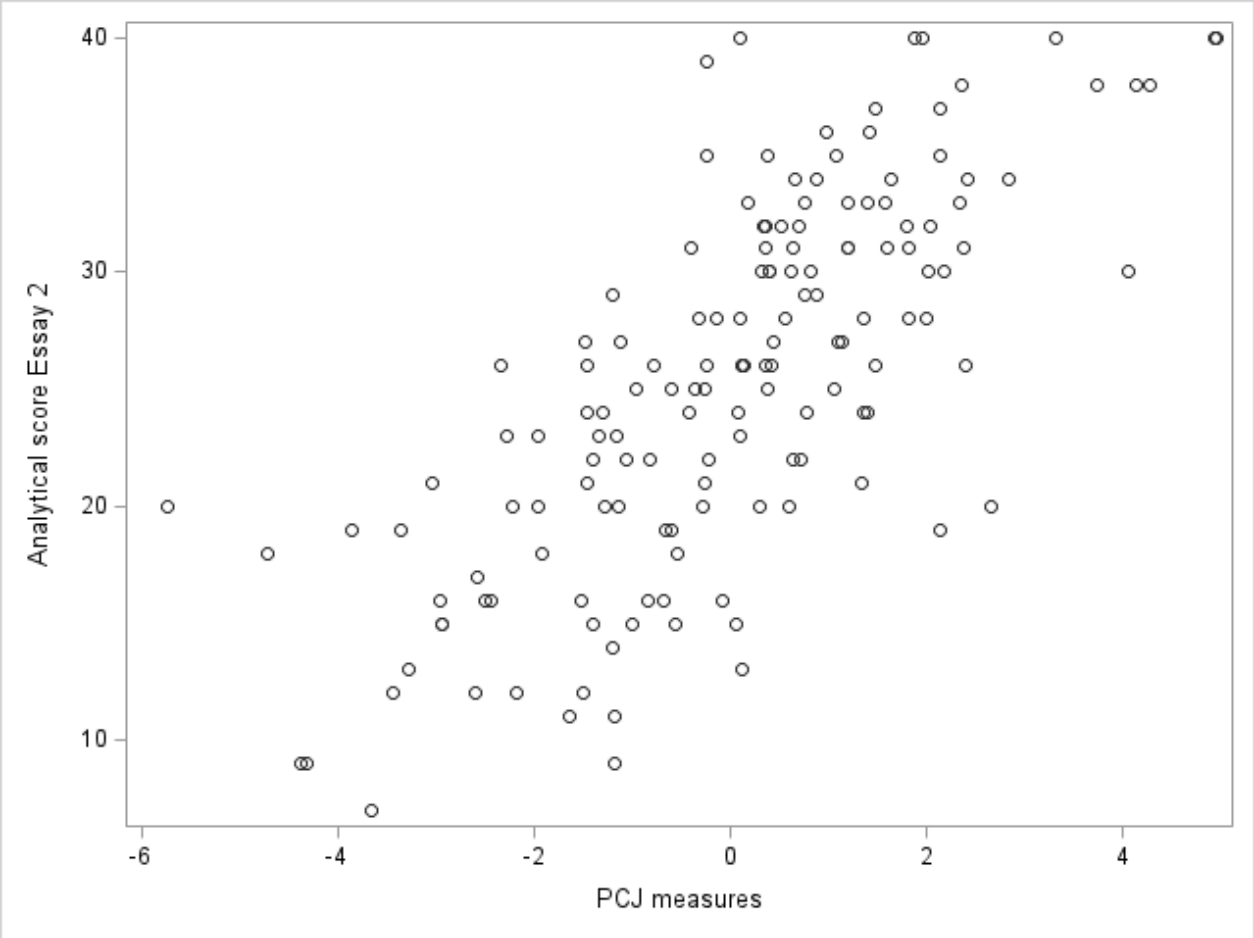

Figure 7. Scatter plot of the relationship between the single marker analytical essay marks for essay 2 (the outcome variable) and the marks generated using PCJ for essay 1 (the predictor variable).

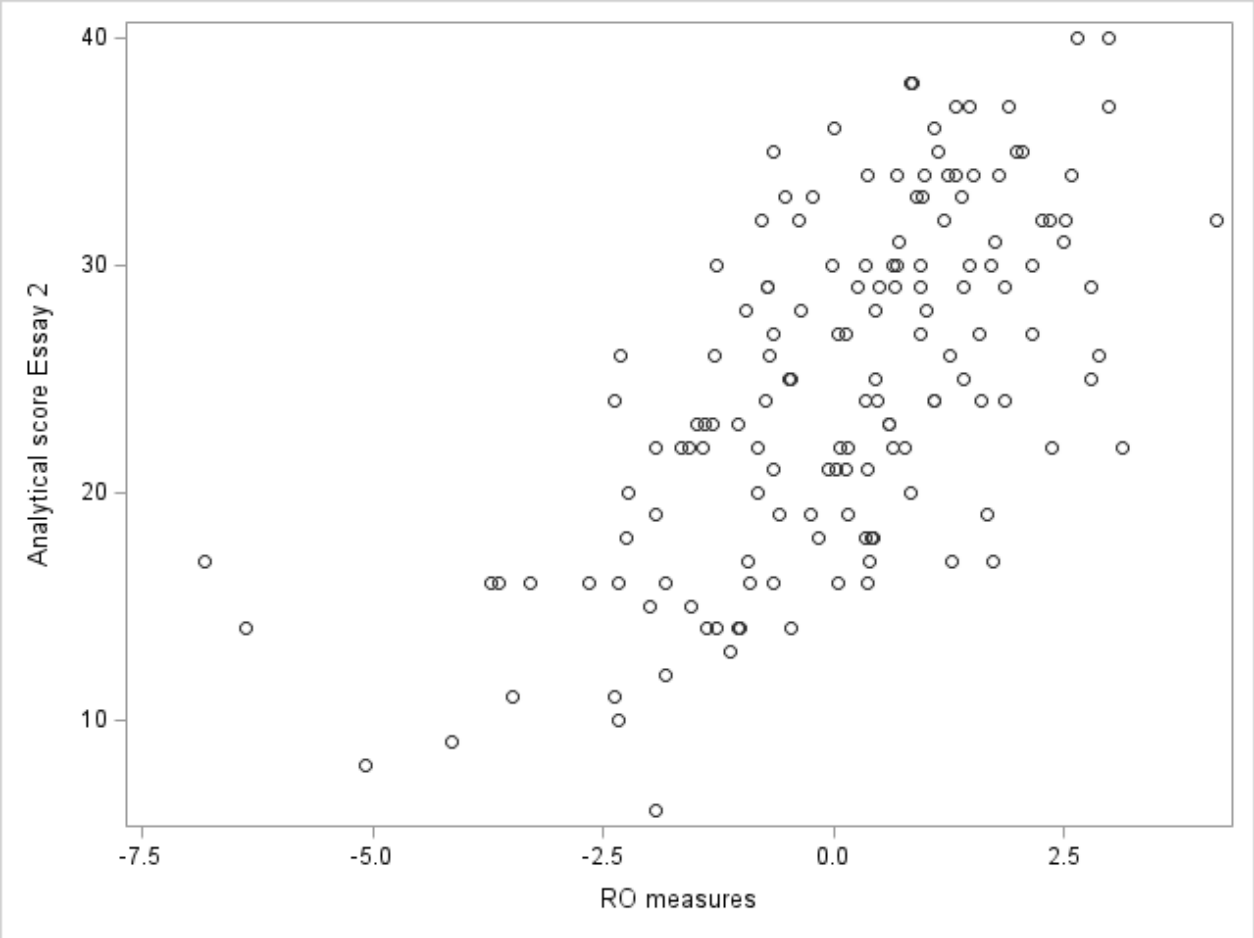

Figure 8. Scatter plot of the relationship between the single marker analytical essay marks for essay 2 (the outcome variable) and the marks generated using RO for essay 1 (the predictor variable).

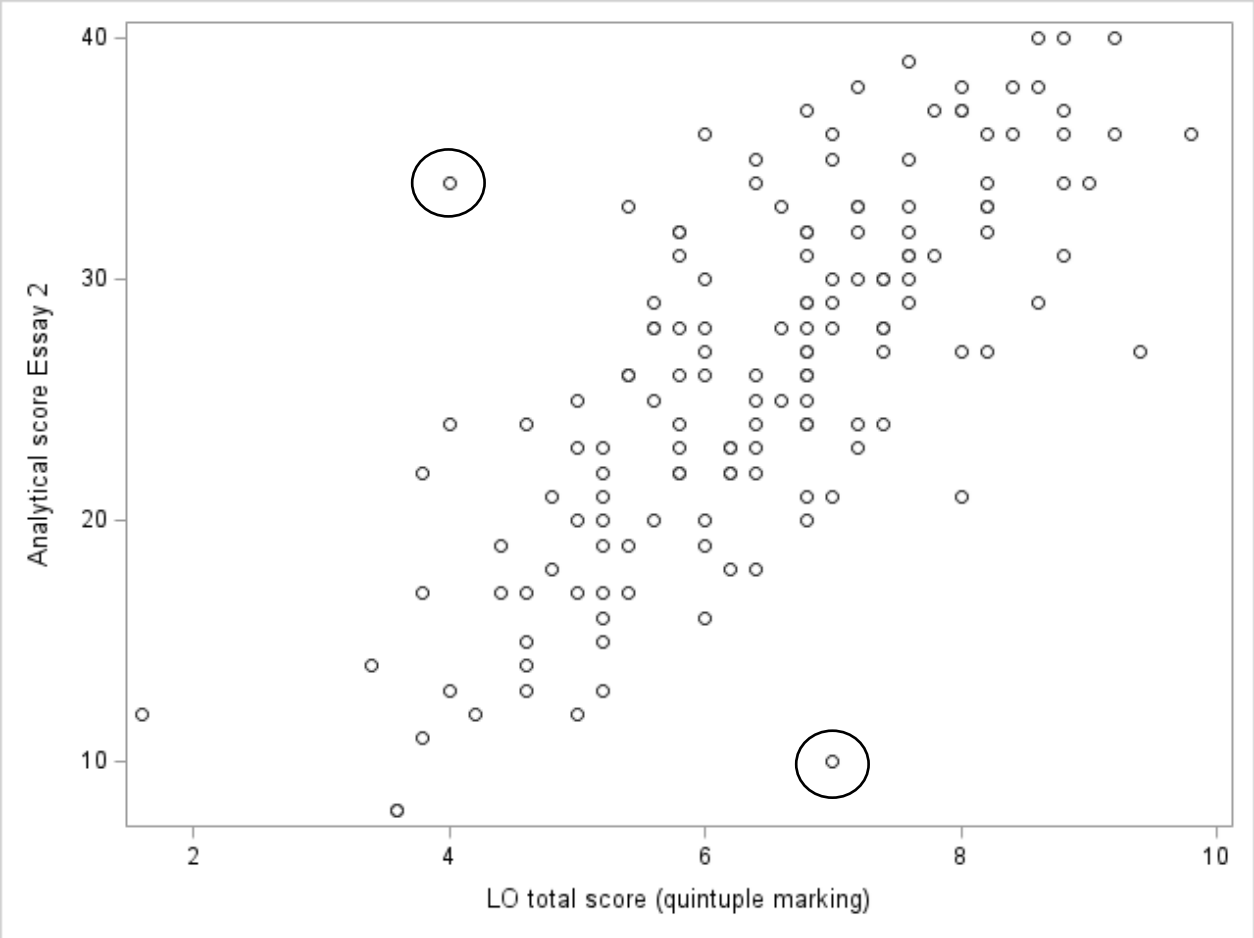

Figure 9. Scatter plot of the relationship between the single marker analytical marks for essay 2 (the outcome variable) and the marks generated using LO quintuple marking for essay 1 (the predictor variable). Two potential outliers with student residuals greater than three are shown in the large circles.

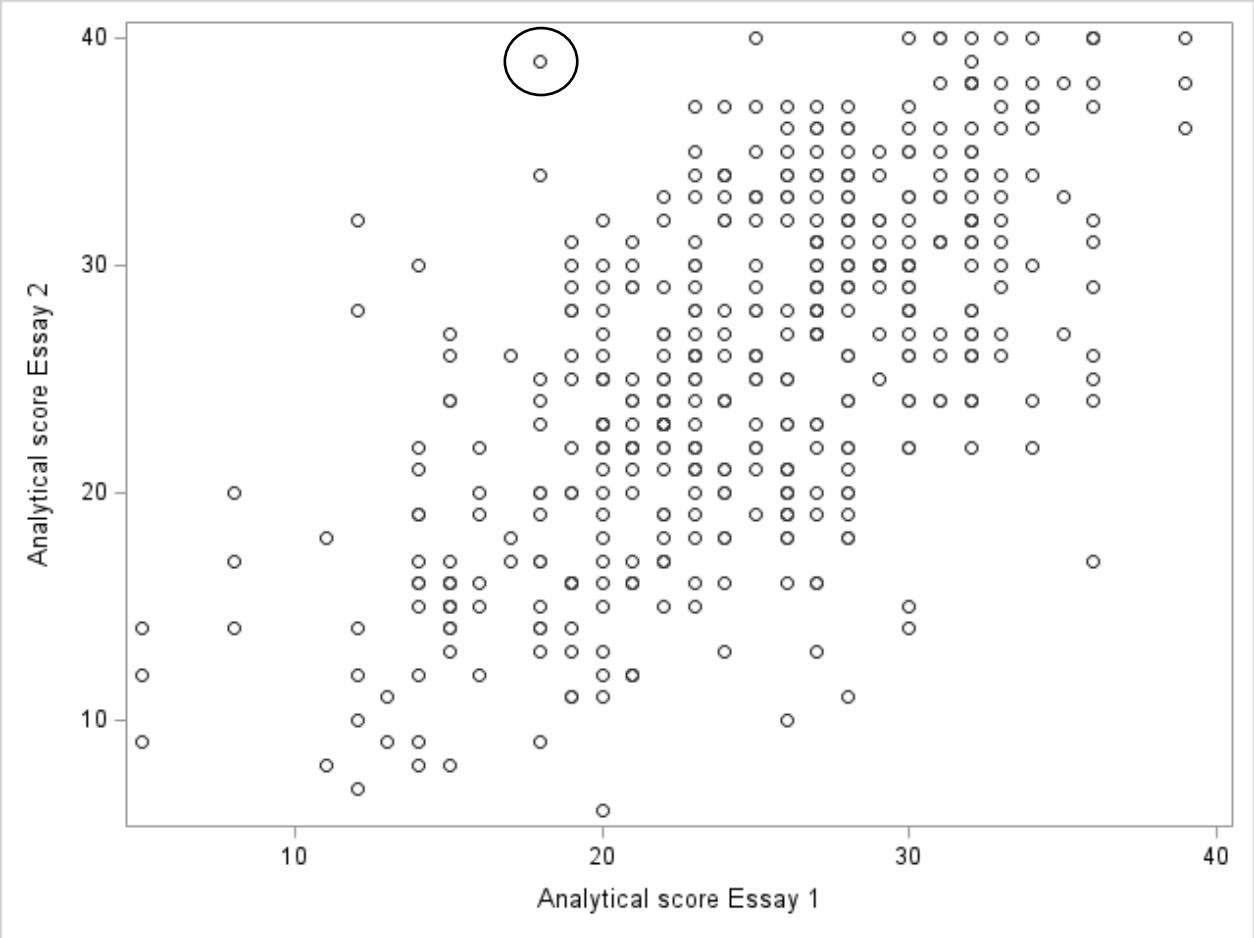

Figure 10. Scatter plot of the relationship between the essay marks for essay 1 (the predictor variable) and essay 2 (the outcome variable) from analytical marking (from the actual marking in 2019). A potential outlier with student residual greater than three is shown in the large circle.

Appendix M – Judging time for single marker analytical marking

To compare any of our results to single marker analytical marking, it was first necessary to have an estimate of how quickly this marking was completed. To do this, we analysed data regarding the amount of time it took to mark essays from the June 2019 examinations. Previous research data established that the mean time taken to mark the whole examination paper was roughly 19 minutes. However, our interest was not in the time taken to mark a full paper but instead in the amount of time required to mark a response to one essay question only. Given that the essay question was worth a total of 40 marks out of the 80 available for the paper, a simple estimate of the required time to mark this question might be nine minutes – that is, roughly half the (robust) mean time for the full paper.

To further verify this estimate, we explored the difference in the amount of time taken to mark each examination paper depending upon whether the essay question had been left blank. We used robust linear regression to estimate the relationship between the necessity to mark the essay (because the essay had been completed) and the amount of time it took to mark each paper. The amount of additional time required when the essay was answered provides a reasonable estimate of how much time it takes to mark this essay on its own. At worst, it provides a lower bound for the estimated amount of time for marking as markers may require a small amount of time to verify for themselves that the response was definitely blank. The regression analysis also accounted for the total marks on the rest of the examination paper (as performance is conceivably related to how long marking takes) and whether each essay was marked by a senior marker, rather than a regular marker, as this could also influence marking speeds. Students who answered the alternative essay option were removed from the analysis as not being relevant. Robust regression was used to enable us to automatically handle outliers in the analysis. A total of 5,431 examination papers were used in the analysis of which 180 did not include any answer to the essay question.

The results of analysis are in Table 16. As can be seen, papers with higher marks tended to take longer to mark (perhaps due to having longer responses on average). In addition, senior markers tended to mark more quickly than regular examiners (although the difference was not statistically significant). However, of most interest was the fact the papers requiring the essay question to be marked took 8.9 minutes longer to mark than similar papers where the essay was blank. These results are consistent with the simple estimate earlier suggesting that, using analytical marking, it took markers nine minutes to mark each essay.

Table 16. Results of robust linear regression.

Independent variable	**Coefficient (minutes extra for marking for a change of one in the independent variable)**	**Standard Error**
Intercept	3.86	0.84
Essay question not missing	8.92	0.88
Senior marker	-0.27	0.35
Total mark on rest of paper	0.33	0.02

Appendix N – Significance tests for predictive value

Table 17 provides significance tests of whether correlations of scores from essay 1 (the predictor variable) for LO, PCJ and RO with scores on essay 2 (the outcome variable) are higher than the correlations between single marker analytical marking and scores on essay 2 (the outcome variable) for the same set of essays. Significant differences are highlighted in bold.

Table 17. Spearman correlation coefficients entered into the online calculator[22] and the results of the significance tests.

Method	Code	Variables	r_s	n	Result
LO Double	12	Analytical and Essay 2	.63	149	*p* = .08
	13	LO and Essay 2	.68	149	z = -1.41
	23	LO and Analytical	.77	150	
LO Triple	12	Analytical and Essay 2	.63	149	***p* = .009**
	13	LO and Essay 2	.71	149	z = -2.36
	23	LO and Analytical	.81	150	
LO Quadruple	12	Analytical and Essay 2	.63	149	***p* = .002**
	13	LO and Essay 2	.73	149	z = -2.89
	23	LO and Analytical	.82	150	
LO Quintuple	12	Analytical and Essay 2	.63	149	***p* = .001**
	13	LO and Essay 2	.73	149	z = -3.06
	23	LO and Analytical	.83	150	
PCJ (18.67 judgements per essay)	12	Analytical and Essay 2	.64	150	***p* = .004**
	13	PCJ and Essay 2	.74	150	z = 2.70
	23	PCJ and Analytical	.76	150	
RO (each essay in 7.5 packs)	12	Analytical and Essay 2	.60	149	*p* = .208
	13	RO and Essay 2	.64	149	z = -0.81
	23	RO and Analytical	.77	149	

Note: the predictive values shown in this table were rounded to two decimal places, but all available decimal places (up to 7) were used for the online calculator.

The following formulae were used to test for the significance of differences in predictive values between different methods trialled with different sets of essays. The results of applying these formulas are shown in Table 18. As can be seen, based on results using the maximum numbers of judgements per essay, no significant differences in predictive value were identified.

$$Z = .5 * \ln\left(\frac{1 + rho}{1 - rho}\right)$$

$$SE = \sqrt{\left(\frac{1.06}{n1 - 3} + \frac{1.06}{n2 - 3}\right)}$$

[22] https://www.psychometrica.de/correlation.html#dependent.

$$ztest = \frac{Z1 - Z2}{SE}$$

Table 18. Results of the Fisher's r to Z transformation to compare the predictive value of RO, PCJ and LO.

	PCJ / RO	**RO / LO quintuple**	**PCJ / LO quintuple**	**PCJ / LO double**
Z (1)	0.96	0.75	0.96	0.96
Z (2)	0.75	0.93	0.93	0.83
SE	0.12	0.12	0.12	0.12
z-test	1.70	-1.49	0.20	1.08

Note: PCJ (n=150), RO (n=149), and LO (n=149). The calculations for PCJ and RO are for the full dataset, i.e., 18.67 judgements per essay for PCJ and each essay in 7.5 packs for RO. The critical value for significance was set at +-1.96 (alpha = .05).