RESEARCH AND ANALYSIS

Improving awarding: 2018/2019 pilots

Milja Curcin, Emma Howard, Kate Sully and Beth Black

ofqual

Table of Contents

List of tables

List of figures

Research in brief

This report looks at some comparative judgement (CJ) methods for capturing expert judgement in the context of standard maintaining ('awarding') in order to derive judgementally recommended grade boundaries. CJ in this context entails multiple judges making comparisons between different scripts (within and between different years, i.e. examination sessions) on the basis of script quality. This allows us to construct a single quality scale ('measure') for scripts across 2 years. Via this common scale of script quality, it is possible to map the known grade boundary marks from a previous year to the equivalent location (script quality) on the mark scale for the current year in order to establish the current year grade boundaries.

Here we report on the findings of a number of pilots in which we trialled different CJ methods – paired comparative judgement and rank ordering – involving a range of expert judges and varying other design features such as the number of judgements per script and types of judges involved. The technical aspects of the pilots as well as the plausibility of the judgementally recommended boundaries are evaluated.

Overall, the results suggest that CJ methods are very promising for capturing expert judgement for the purpose of standard maintaining. The totality of the pilots indicate that pooling a sufficiently large number of judgements over most of the mark range can give reliable outcomes and potentially increase the validity of expert judgement in standard maintaining.

Further consideration needs to be given to the merits of different designs in operational contexts, and the relative weight such methods might carry in relation to statistical indicators used in standard maintaining.

Executive Summary

Background and motivation

Maintaining standards between examination sessions should ensure that students with the same level of attainment get equivalent grades on exams from different sessions. In GCSE and A level qualifications standard maintaining currently primarily relies on a method often referred to as the comparable outcomes approach (Cresswell, 2003; Bramley and Vidal Rodeiro, 2014). This method involves statistical predictions that model the relationship between prior attainment and outcomes in a reference year, then apply this relationship to the current cohort (Taylor and Opposs, 2018). When used as the sole method for maintaining standards, it operationalises an assumption that cohorts of similar ability should have similar pass rates and similar grade profiles in different examination sessions.

This value-added approach takes account of changes in the prior attainment of the cohort, but makes it difficult to recognise changes of cohort performance due to changes in the overall quality of teaching and learning over time, for instance as a result of explicit changes in the curriculum or teaching methods to drive performance improvements. To mitigate for this, in addition to statistical recommendations for where grade boundaries should be set, student work (scripts) on key statistically recommended boundaries (SRBs) is scrutinised by expert examiners to help ensure that the new grade boundary performance standard reflects that from previous sessions as well as performance changes.

However, some limitations in the way expert judgement is currently captured means that it might not be as strong a source of evidence as it could be in regard to maintenance of standards. Notably, expert recommendations are quite explicitly guided by statistical recommendations since generally only a narrow range of score points around the statistically recommended cut scores is considered by experts. Therefore, evidence from expert judgement cannot be treated as an independent source of evidence. Furthermore, it is based on only a relatively small number of judgements within a relatively narrow mark range, limiting the reliability of expert judgement.

Prior research suggests that comparative judgement (CJ) may be a promising alternative method for maximising the reliability of expert judgment about script quality. In this report, we present the results from piloting several comparative judgement methods for capturing expert judgement in awarding. The pilots were conducted in specifications with at least 3 years of awarding, in 4 different qualifications (GCSE media studies, AS English literature, AS psychology and GCSE English language) across 4 exam boards. Where possible, the pilots were conducted during live examination sessions, prior to awarding, while some were conducted outside of live marking and awarding.

In CJ methods, series of comparisons of candidate scripts from different exam sessions are made by several judges (within and between different years, i.e. examination sessions) on the basis of script quality. This allows us to construct a single quality scale (‘measure’) for scripts across 2 years. Via this common scale of script quality, it is possible to map the known grade boundary marks from the

previous year (Y1) to the equivalent location (script quality) on the mark scale for the current year (Y2) in order to establish the current year grade boundaries.

Methods

The methods trialled were rank ordering (RO), online paired CJ (PCJ) and a 'pinpointing' approach (a hybrid of RO and PCJ), each using 6 to 20 expert judges, involving 10-36 judgements per script. In addition, a 'crowdsourcing' online PCJ was conducted, using 40 teachers as judges, each of whom made a small number of judgements. The pilots also varied in terms of the range of mark points included (around 50% or around 70% of mark points from the effective mark range in each session).

Main results

The results of both the RO and the PCJ larger-scale pilots in English language (with 25 comparisons per script and 70% of the mark range) consistently show that the comparative judgement exercises succeeded in producing plausible script quality scales (with scale separation reliability (SSR) of 0.9 or higher) and high levels of agreement between original test score scales and script quality measure scales (correlations of 0.9 or higher). Furthermore, the slightly smaller-scale pilots based on about 50% of the mark range and 20 comparisons per script, appeared to work equally well. This would suggest some scope for streamlining data collection in operational settings.

In the PCJ pilots conducted with only 10-12 comparisons per script, the SSRs ranged from 0.7 to 0.8. Mark-measure correlations tended to vary between 0.6 and 0.8 in these pilots. While most of the grade boundary estimates from these smaller pilots were still plausible, we suggest caution in a few cases in English literature pilots where either mark-measure correlations or the SSRs, alongside very wide confidence intervals, resulted from the pilots. The RO pilots in media studies and psychology, where 20-36 comparisons were collected per script, produced SSRs of 0.9 or higher, mark-measure correlations of 0.75 or higher and largely plausible grade boundary estimates. Reassuringly, the results of different CJ methods (RO and PCJ) carried out on Psychology and English language largely cross-validated each other even where smaller number of comparisons per script were collected, producing very similar grade boundaries, while the script quality measures from these pilots were mostly highly correlated.

English literature RO and teacher PCJ pilots in some cases resulted in fairly low mark-measure correlations and/or SSRs lower than 0.7. The average of 20 comparisons per script did not seem to help mark-measure correlations despite reasonable SSRs in these exercises. Furthermore, the correlation of the measures produced in examiner vs. teacher PCJ exercise in English literature correlated less well than where exercises were replicated with different methods in English language and Psychology. These results in English literature may have resulted from an interplay of smaller-scale exercises, relatively short and possibly less discriminating test score scales, and, to some extent, possible incongruence between aspects of the original mark scheme and the features considered in holistic judging.

Unlike most of the abovementioned pilots, when we trialled the pinpointing approach, this failed to result in convincing script quality scales, and produced some

implausible grade boundary estimates. We would suggest that this approach, with its focus on a narrow range of mark points around the grade boundaries, may not be the optimal way of maximising judge and scale reliability.

Further evaluation and considerations

While in most cases plausible script quality scales were associated with grade boundaries that were largely congruent with the Y2 operational boundaries, in some cases, they were associated with the boundaries that were fairly discrepant from the Y2 operational ones. In general, and in the latter case in particular, it would be necessary to consider a range of available sources of evidence and give appropriate weight to these sources in deciding on the most likely appropriate grade boundaries.

Examining the patterns of differences between pilot and operational boundaries across all the pilots suggests that there was little evidence of consistent positive or negative differences compared to Y1 or Y2 operational grade boundaries. Thus, it seems unlikely that the results were consistently affected either by idiosyncrasies of the RO or the PCJ methods, or by deliberate 'gaming' by the judges (e.g., always 'preferring' the scripts from the more recent session, which might lead to higher outcomes for candidates in that session). Furthermore, the comparison of consistency levels in within- versus between-session judgements did not reveal any worrying differences that would suggest that the between-session comparisons may have been less consistent or degrading the measurement process, at least in the English language pilots. This would be an important aspect to monitor in all CJ exercises. Reassuringly, the qualitative analysis of the strategies and response features that were reported by the judges suggests that the judges were using mostly valid strategies and response features when making their judgements, which accord with those identified in prior research. Most of the response features identified were also present in relevant mark schemes.

We also looked at the effects of judge expertise (e.g., ordinary vs. senior examiner) on CJ outcomes, concluding that there does not appear to be a tangible and consistent effect of this. This suggests, alongside other research, that ordinary examiners can participate in these exercises without compromising our confidence in the outcomes. This could potentially help organise CJ exercises during live marking, as it could free up senior examiners to deal with their other obligations during this period.

Regarding using the confidence intervals derived from bootstrapping to quantify likely variability in CJ outcomes, we have argued that traditional confidence intervals based on +/-2SD around the mean might be too stringent in this context. We suggested that middle 50% IQRs might be appropriate, especially within the constraints of exercises with similar design parameters and judge expertise year on year. This may also be considered appropriate given the apparent robustness of these methods to a range of design and other manipulations, as well as replicability of the results in different contexts and with different judges.

Furthermore, considering the bootstrapping results, it is clear that there is more potential variability at GCSE grade 1 in particular, but also at grade 9, and to some extent at AS level grade E. This is a familiar effect with respect to more extreme scores. It would be important to consider how to overcome this challenge in judgemental exercises.

With respect to optimal number of mark points and scripts to include in CJ exercises (for instance, in our case, 50% vs. 70% of mark points), it should be noted that, ideally, the sample of scripts used in CJ exercises should be in some way representative of the full set of scripts from the relevant examination (cf. Benton, 2019). Reducing the number of mark points included in CJ exercises would potentially reduce the representativeness of the sample further, potentially leading to grade boundary outcomes that would not be representative of the outcomes that would have been obtained if the full set of mark points and scripts was judged. Furthermore, where smaller number of scripts is used, it would be important to consider the implications this would have for bootstrapping analysis, as its results may be less valid (for example, may appear to overestimate the variability in the outcomes) when there is a small number of objects in the sampling pool. While the number of scripts to be included in CJ exercises may be limited by practical considerations in operational contexts, the precise impact of different sample sizes and profiles requires further research.

Regarding judges' ability to compensate for differences in paper difficulty between sessions in their script quality judgements, which is one of the assumptions of the CJ methods when used for standard maintaining, there is some indication that the judges may be able to do this, as they referred to reasonable techniques of doing so and aspects of question demands when accounting for this in their survey responses. However, it is not easy to see clear patterns of alignment between judges' initial views of empirical paper difficulty and the corresponding pilot outcome. For example, a large number of judges thought that the papers from different sessions were 'similar', irrespective of the final outcome. Arguably, however, most of the grade boundary differences between sessions were indeed very small, and possibly justify a view that papers where boundaries between sessions differ by one or two marks can reasonably be described as similar. Additionally, for some of the papers in this pilot, some of the approaches to marking might change between years, for instance, there may be changes in leniency or severity to awarding top level mark bands. This means that there is not a straightforward link between strict paper difficulty (paper difficulty as an aggregate of the tasks, not the marking approach) and the relationship between the 2 mark scales. All of this is an area that needs more exploration. It would, however, seem important to be realistic about the level to which judges can reasonably be expected to be able to account for differences in empirical test difficulty to the extent a statistical equating method based on large quantities of data could. This probably needs to be recognised as an unavoidable source of error and a shortcoming of all approaches to standard maintaining that do not rely on pre-testing test items or test forms routinely, which could enable robust statistical equating methods to be used.

Conclusions

While further consideration needs to be given to the merits of different CJ methods and specific designs in operational contexts, overall, the results of our pilots suggest that CJ methods are very promising for capturing expert judgement for the purpose of standard maintaining. The totality of the pilots indicate that pooling a sufficiently large number of judgements over most of the effective test score scale can increase the reliability of the outcome of expert judgement, potentially increase the validity of expert judgement in standard maintaining and thus increase our confidence in expert judgement recommendations. The fact that CJ methods are implemented

independently of statistical grade boundary recommendations and knowledge of original script marks helps to preserve judgemental evidence as an independent source of evidence that could be attributed its own weight appropriate to the specific context of use.

Some pilots were designed in such a way as to facilitate achieving SSRs of around 0.9. However, lower levels of reliability might be considered appropriate, and therefore smaller-scale exercises, which might still be sufficiently robust, may be reasonably attempted in some contexts. Decisions about the scale of CJ exercises might also need to be driven by the intended weight that might be given to judgemental evidence in each case. Where more weight might be placed on the judgemental outcomes (for instance, where there is less confidence in the statistical outcomes for whatever reason) it might be reasonable to collect judgemental data on a larger scale.

It would also seem important to continue investigating suitability of different criteria for evaluating the comparative judgement methods, including appropriate confidence intervals for grade boundary estimate precision. Evaluation criteria for judgemental methods might to some extent depend on the way we conceptualise their place in awarding. While these methods certainly go a long way towards enhancing the reliability of expert judgement and increasing our confidence in its recommendations, it may still be inappropriate to attempt to evaluate them according to stringent criteria that may be applicable for purely statistical methods of equating.

Introduction

Because tests from different session can vary in difficulty, standard maintaining focuses on establishing equivalent marks and grade boundaries on later versions of a test which carry over performance standards from an earlier version of the same test, adjusting for test difficulty differences. However, in the context of GCSE and A level qualifications, traditional statistical methods used in some other jurisdictions (see Kolen & Brennan, 2004) cannot be used to maintain standards as there is no pre-testing of items or tests on representative samples of examinees, where common items or common candidates would reflect changes in test difficulty or cohort ability. Pre-testing or reusing items for this purpose is impractical in the assessment model for GCSE and A level given the high stakes nature of each exam series and security considerations.

Maintaining standards from one examination session to the next (awarding) in GCSEs and A levels currently primarily relies on a method often referred to as the comparable outcomes approach (Cresswell, 2003; Bramley and Vidal Rodeiro, 2014). When used as the sole method for maintaining standards, it tends to be used to operationalise the assumption that cohorts of similar ability should have similar pass rates in different examination sessions. In other words, grade distributions in consecutive examination sessions, or in the current vs. a specified reference session, should not change if the relevant cohorts are of similar ability.

When interpreted as a principle, rather than simply as a method, the comparable outcomes approach prioritises comparable outcomes over comparable performance year on year (which is contrary to traditional grade awarding practices). One good reason for this is that it protects students taking their assessments in the first year of a new qualification, when teachers and students are less familiar with the assessment and performance is likely to dip (Taylor and Opposs, 2018; Cuff, Meadows and Black, 2018). However, in ‘steady-state’ qualifications (i.e. after the first few years of examinations), protecting students in this way arguably becomes less relevant, while recognising any genuine changes in performance due to overall changes in teaching and learning, becomes important.

The comparable outcomes approach derives prior attainment based predictions, which map the relationship between prior attainment (i.e. ability) and GCSE or A level outcomes for students taking each subject in a reference year, and use this relationship to predict the outcomes for the current cohort of students based on their prior attainment (see Ofqual, 2019a, 2019b for further details). Thus, if prior attainment (ability) of the cohort remains similar in different examination sessions, the outcomes would be expected to be similar too (Taylor and Opposs, 2018). At A level, the measure of prior attainment is the mean GCSE score. At GCSE, the measure is based on Key Stage 2 performance.

As Bramley and Vidal Rodeiro (ibid.) point out, when used (in effect) as the sole method for maintaining standards in steady-state qualifications, the result of the comparable outcomes approach would seem to support inferences of the kind ‘Students with a grade A this year have (on average) the same level of prior attainment as students with a grade A last year (or from a different board)’. However, according to these authors, the majority of stakeholders’ interpretations of grades in different examination sessions are more likely to be along the lines that a student with the same level of attainment should get the same grade on different exams in

the same subject – either over time within board or across boards. Arguably, unlike the former interpretation, the latter interpretation is akin to those enabled by traditional equating methods, where the equated marks on the current test can be seen as interchangeable with those from a previous test (Bramley and Vidal Rodeiro, ibid., cf. Kolen & Brennan, 2004).

In order for such inferences to be supported, it would be necessary for the method by which the grade boundaries are established in each examination session to be appropriate for such interpretations and based on defensible assumptions. The situation in which comparable outcomes approach is used requires a number of strong assumptions, which would normally be problematic in traditional equating, about the nature of the tests being equated, the nature of the measures of prior attainment which crucially feed into the prediction matrices, etc. Bramley and Vidal Rodeiro (ibid.) argue that when comparable outcomes method is used as an equating method (to link exam-related attainment, rather than prior attainment, or ability), it would be necessary to evaluate the plausibility of any assumptions that must hold for the statistical equating approach to give accurate results, especially where the statistical approach is used as the dominant source of evidence about where to set the grade boundaries. Furthermore, where the assumptions can be shown not to hold, or to be implausible, it would be important to understand what implications this has on how grade boundaries should be set (Bramley and Vidal Rodeiro, ibid.).

Following Newton (2011), Bramley and Vidal Rodeiro (ibid.) characterised traditional grade awarding practices as constituting a relatively relaxed conception of standard maintenance and equating, which allows for statistical input to the setting of grade boundaries, but also for the consideration of other evidence such as expert judgment of the quality of work produced. Inclusion of other sources of evidence could lead to setting boundaries that deviate from those derived statistically, on the basis of comparable outcomes.

Indeed, Ofqual's now retired Code of Practice (Ofqual, 2011) stated that the decisions on where to locate the grade boundaries should draw on a number of sources of evidence, rather than just the results of the statistical methods (Code of Practice section 6.13), although the statistical evidence is mostly dominant (cf. Ofqual 2019a, b). This is probably in recognition of the fact that it is not always straightforward to be certain that all of the assumptions required by the statistical models have been met. This means that statistical evidence cannot provide absolute certainty regarding the abovementioned interpretation of the meaning of the same grades from different examination sessions. Furthermore, the statistical techniques used to model the value added relationship between 2 examination sessions in the comparable outcomes approach cannot factor in potential performance changes when determining current grade boundaries.

Awarding, therefore, also involves consideration of expert judgement about exam difficulty and the quality of candidate work at statistically recommended grade boundaries (SRBs), which is intended to provide another source of evidence that performances on different examinations represent the same performance standard at key grades. Currently, examiner judgement represents an important check of the plausibility of the statistical recommendations (see Ofqual, 2019a, b) rather than a distinct source of evidence regarding equivalent grade boundaries. Stringer's (2012) research suggest potential risks of such position of expert judgment in the current

system, describing a situation in which an awarding committee failed to spot when the statistical predictions for a particular paper were calculated incorrectly leading to biased results (initially), while another committee was able to spot and address an error in statistical predictions. As Benton and Bramley (2015) argue, these examples indicate some expectation of examiners not to depart too far from statistical predictions, rather than their inability to reliably use their judgement when allowed to exercise this free from the influence of statistics.

There are some limitations in the way expert judgement is currently captured meaning that it might not be as strong a source of evidence as it could be in regards to standard maintaining. This means it is likely to be difficult to make a strong case for moving away from the statistically recommended boundary based on expert judgment alone. This is not least because expert judgement is informed by the statistical recommendations in the first place, the judgements within a relatively small range of marks around key grade boundaries are difficult to make reliably, and any expert recommendation is based on a very small number of judgements.

Prior research suggests that comparative judgement (CJ) may be a promising alternative method for capturing and maximising the reliability of expert judgment about script quality. In this report, we present the results from piloting different CJ methods: rank ordering (RO), online paired CJ (PCJ) and a 'pinpointing' approach (a hybrid of RO and PCJ). The pilots were conducted in specifications with at least 3 years' of awarding, in 4 different qualifications across 4 exam boards. Where possible, the pilots were conducted during live examination sessions and prior to awarding, while some were conducted outside of live marking and awarding.

Current procedures for capturing expert judgement of script quality in awarding

Expert judgement informs the setting of the 'key' grade boundaries. These are A and E in AS and A level and 7, 4 and 1 in GCSEs (A, C and F in unreformed GCSEs). The remaining grade boundaries are determined arithmetically.

Taylor and Opposs (2018) provide a description of how expert judgement is currently captured in awarding. The main source of judgemental evidence is script scrutiny. As part of an awarding meeting, usually 4 to 6 senior examiners from the relevant specification are presented with exam scripts on a range of marks (typically 3 to 5 marks) around the statistically recommended grade boundary. They must independently decide whether each exam script in this range is worthy of the grade under consideration or not. In doing this, examiners are required to refer to archive scripts on the grade boundary marks from previous years and statistical evidence showing the performance of individual questions on each exam paper.

The examiners' judgements are recorded on a 'tick chart', as shown in Figure 1. A tick means that a committee member thinks that the work is worthy of the higher grade of the boundary pair (for instance, A/B), a cross means that they do not, and a question mark means that they have some doubts. Based on the balance of ticks and crosses, the chair of examiners specifies a 'zone of uncertainty' – illustrated here in grey. This is the zone within which the judgmental evidence suggests that the grade boundary should lie.

Table 1 *Awarding committee judgements of script evidence*[1]

Mark	Chair of Examiners	Chief Examiner	Principal Examiner A	Principal Examiner B	Principal Examiner C	Principal Moderator
54	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✓✓✓
53	✓✓✓	✓✓✓	✓✓?	✓?✓	✓?✓	✓✓✓
52	✗?✗	✓✗✗	✓✗✓	✓✗✓	✓✗✗	✓✗✗
51	✓✗✗	✓✗✗	✗✗?	✗✗✗	✗✗✗	✗✗✗
50	✗✗	✗✗	✗✗	✗✗	✗✗	✗✗

This description shows the typical scale of judgemental exercises currently conducted as part of awarding. Several studies have investigated the reliability of expert judgements collected on a similar scale, concluding that judgement reliability is too low to support basing awarding decisions on expert judgement alone or using it with confidence to move away from the statistically recommended grade boundary (Baird and Dhillon, 2005; Forster, 2005 (cited in Benton and Elliott, 2016); Good and Cresswell, 1988; Cresswell, 1997; etc.). However, as and Benton and Elliott (ibid.) show (cf. Benton and Bramley, 2015), low reliability of expert judgement demonstrated in those studies may be increased if using methods such as comparative judgement, which typically collect expert judgement on a much larger scale and using relative rather than absolute judgements.

Collecting expert judgement on a larger scale for standard maintaining

Benton and Elliott (ibid.; cf. Benton and Gallacher, 2018: 22) demonstrate through simulation work that low reliability of the judgements made in the above-mentioned studies is likely due to small numbers of judges, scripts and judgements used. They show that by combining judgements across a larger number of examiners judging a larger number of scripts, and using statistical models to iron out differences in the severity of different judges, it is possible to increase the reliability of the judgemental process and its outcomes (see also Bramley and Vitello, 2019).

Further support for the argument that expert judgements of reasonably high reliability can indeed be obtained subject to sufficiently large-scale judging design is provided in Verhavert et al. (2019). They conducted a meta-analysis of the results of 49 CJ assessments, where paired comparisons were used instead of traditional marking/rating of a range of performance assessments. The results show that between 10 and 14 comparisons per performance are needed for reliability levels of 0.70, which might be appropriate for low stakes situations such as formative assessments; 26 to 37 comparisons per performance (20 to 35 comparisons per performance for expert examiners) are needed for a reliability of 0.90, arguably more appropriate for high stakes decisions in terms of students' individual location in the resulting rank order of ability/performance quality.

[1] Reproduced from Taylor and Oppos (2018).

Expert judgement collection can be implemented on a larger scale via CJ methods, through either RO or paired comparisons. CJ methods have been investigated as a potential alternative source of information for standard maintaining, i.e. 'test equating by expert judgement' (Bramley, 2005; Black and Bramley, 2008; Raikes, Scorey and Shiell, 2008; Black and Gill, 2008; Bramley and Gill, 2010; Gill and Bramley, 2013). In addition, they have been investigated and used as an alternative to traditional marking (Pollitt, 2004, 2012; Kimbell et al., 2009; Jones and Inglis, 2015; Jones, Swan and Pollitt, 2014; Steedle and Ferarra, 2016; Heldsinger and Humphry, 2010; Verhavert et al., 2019), and in inter-board comparability studies (Bramley et al, 1998; Elliott and Greatorex, 2002; Pollitt and Elliott, 2003; Jones et al., 2004; Jones et al., 2016).

The theory underlying CJ methods is Thurstone's law of comparative judgement (Thurstone, 1927). Given that the judgements collected through CJ methods are subsequently analysed using the Rasch model, which allows for non-randomly missing data, it is possible to create judgment allocation designs that are sufficiently sparse to be feasibly implemented in practice, while being sufficiently large-scale to result in reasonably precise estimates of the scale of interest (for example, the script quality scale). Bramley (2007: 279) notes that a number of RO studies have shown that in CJ exercises within a judge's allocation of judgements the correlation between perceived quality and mark for scripts from the same test is often low, and sometimes even negative. However, when the results are aggregated over the entire mark range for all judges, the overall correlation between mark and measure of script quality is high (around 0.8 to 0.9), which again suggests that pooling sufficiently large number of judgements can increase the validity of the outcome of expert judgement, and, by extension, our confidence in expert judgement recommendations.

The main theoretical advantages of CJ methods are the experimental elimination of the internal standards of the judges when estimating scale locations, and the fitting of an explicit statistical model that allows investigation of residuals for script and judge misfit, and for various sources of bias (Bramley, 2007; Pollitt and Elliott, 2003). An additional feature of CJ methods is that the judgements elicited through them are arguably psychologically easier to make and more intuitive than absolute judgements of the kind made in current awarding procedures (e.g., Thurstone, 1927, 1931; Laming, 2004; cf. Baird, 2000). There is some debate in the context of examination marking regarding whether the increased reliability of CJ judgements may be solely a virtue of the scale on which they are collected and statistical models used for data processing rather than being related to psychological factors (Benton and Gallacher, 2018). However, in contexts such as awarding, when mark schemes cannot be used, there is some evidence in the literature that holistic comparative judgements may indeed be more accurate than holistic absolute judgements of the kind currently made during awarding, (Gill and Bramley, 2013). This is, however, probably unlikely to be over and above the effect of collecting these judgements on a larger scale in the first place.

CJ methods in standard maintaining

While studies in the context of marking and inter-board comparability tended to use paired comparisons, the studies investigating standard maintaining between different examination sessions via expert judgement tended to use the RO method. In this method, judges are given packs of student examination scripts without marks and annotations. Each pack contains scripts from two or more tests. The packs usually

cover most of the effective mark range. The scripts in each pack vary in the range of marks covered and the degree of overlap of the mark ranges from each test. Each pack contains a unique selection of scripts, but it is necessary to ensure that there are common scripts between the packs in order to link the entire set of scripts to enable common scale estimation. Each judge is given a number of packs covering most of the mark range, where packs containing scripts with higher total marks are usually presented first. The number of times each judge sees the same script is minimised but some repetition is usually necessary.

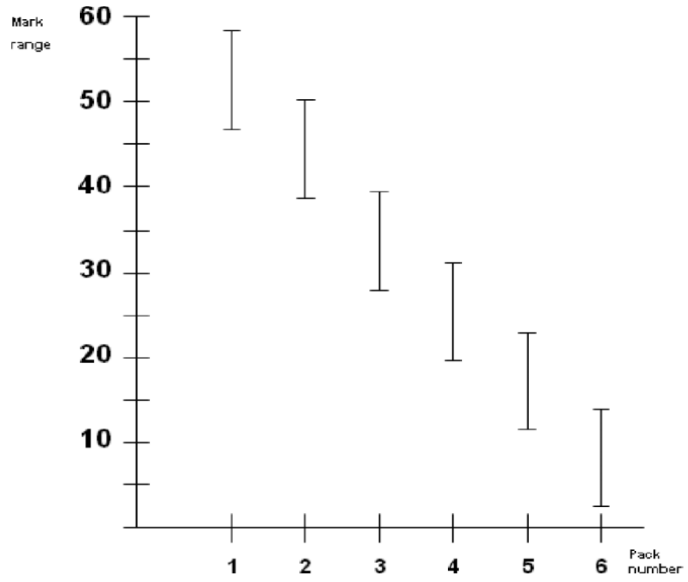


Figure 1 *Typical pack design for RO studies*[2]

The judges rank order the scripts in each pack in terms of quality from best to worst. They are asked to allow for differences in difficulty between the question papers from each examination session.

The rank orders are usually converted into paired comparisons and a single 'perceived script quality' scale across different examination sessions is derived using a Rasch formulation of Thurstone's (1927) paired comparison model (Andrich 1978; see Bramley, 2007).[3] Each script is positioned on this scale in terms of quality, which is related to the probability of it being judged better than another script in a paired comparison. The same model can be used to analyse judgement data from a paired comparisons exercise. The model can be stated as:

$ln[P_{ij} / (1 - P_{ij})] = B_i - B_j$
where P_{ij} = the probability that script *i* beats script *j* in a paired comparison
and B_i = the measure for script *i*
and B_j = the measure for script *j*

The unit of the scale created by the analysis is known as a 'logit' or 'log-odds unit'. The analysis can be carried out using the FACETS software (Linacre 2019) or the

[2] Reproduced from Black and Bramley (2008).
[3] Another approach is to use a model for rank-ordered data, allowing for the constraints imposed by a ranking. This is implemented in Placket-Luce model (Plackett, 1975) as implemented in PlackettLuce R package. The Rasch formulation of the Thurstone model is a special case of this more general model.

sirt package in R, which implements the Bradley-Terry version of the model (Bradley and Terry, 1952).

As Bramley and Gill (ibid.) explain, the RO method (as well as the paired comparisons method) has the same conceptual foundation as the latent trait statistical equating methods. The tests from different examination sessions are assumed to measure the same trait, on which the examinees can be compared in terms of ability (or their scripts in terms of quality). The differences in test scores are assumed to reflect the differences in trait level among examinees. However, because the tests from different sessions could differ in difficulty, the same raw marks on 2 different tests do not necessarily imply the same trait level (ability) – hence the need for equating via the common ability/quality trait. The main assumption of CJ methods used for equating by expert judgement is that expert judges can directly perceive differences in trait location among the scripts being judged, implicitly allowing for differences in difficulty of the test forms between different examination sessions.

In order to equate the tests from 2 sessions using the script quality scale, this scale is regressed onto the original test score scale. The regression lines summarise the relationship between mark and measure and allow identification of the cut scores for the current examination session that correspond to the cut score performance quality from the previous session.[4] The equivalent mark for each grade boundary of interest (or any other score point) is determined by inserting the previous session cut-score into its corresponding equation to determine the corresponding measure, then inserting that measure into equation for the current session in order to determine the equivalent current session cut-score. In figure 2, a mark of 25 on Test A corresponds approximately to a mark of 28 on Test B in terms of the quality of the scripts on those mark points.

[4] Bramley and Gill (ibid.) demonstrate that the outcomes are fairly constant when the method of plotting the best-fit line are varied and conclude that there is not any very convincing reason to shift from the $Y|X$ regression of test score on measure which has been used in rank-ordering studies to date. However, recent work by Benton (2019) demonstrates via simulation that there may be good reasons for preferring measure on mark regression, although the differences in substantive outcomes may be small – half to one mark difference in grade boundary estimates. This is an issue for further research.

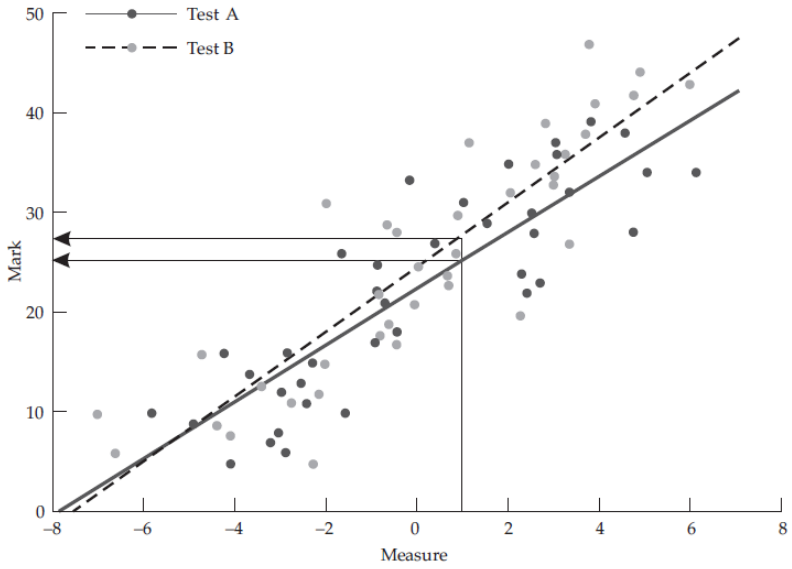


Figure 2 *Example of test equating by expert judgement using the RO method*[5]

Given that these expert judgements are collected from judges individually, independently of one another, rather than in a group setting where people are likely to be aware of others' judgements, and are not influenced by where the statistically recommended grade boundary should be, this can provide an independent source of evidence, alongside the statistical recommendation. However, as Bramley and Gill (ibid.) point out, in the context of standard maintaining in GCSE and A level examinations there is no obvious and completely uncontroversial right answer regarding the correct grade boundaries in a particular examination session from either the statistical or judgmental methods. It is, therefore, necessary to evaluate both the statistical and judgemental methods with respect to stability and replicability of their outcomes, and consider any available contextual information to inform setting of the most appropriate grade boundaries.

With respect to CJ methods, it is necessary to have confidence that, whichever specific design and methodology is employed to collect expert judgements, the outcome is not going to be overly influenced by its artefacts. Importantly, as Bramley and Gill (ibid.) point out, the RO method (this also applies to the paired comparisons method) is a 'strong' method of capturing expert judgement, in that it is possible to evaluate the extent to which it has worked in a given situation in at least 2 key ways. It is possible to establish whether:

- a meaningful scale of perceived script quality has been created, based on the statistical properties of the scale, and
- the perceived quality scale is sufficiently correlated with the test score scale

These and other aspects of the CJ methods that can and should be evaluated in each study are described in the next section.

[5] Reproduced from Bramley (2007).

Evaluating the design features and results of CJ methods

Design features of CJ methods

With respect to designing a CJ study, the goal is usually to maximise the number of comparisons of available scripts across available judges and time, while keeping the exercise practically feasible. The judging design needs to ensure sufficient linking between scripts and judges, i.e. that each script is seen by multiple judges, to allow estimation of the single perceived quality scale using a Rasch model.

While the basic principles of a CJ study design are essentially the same, a number of design features have varied across studies. Bramley and Gill (ibid.) list various factors incidental to this method (some of these are also relevant for PCJ):

- the number of scripts to use
- the criteria for selecting and/or excluding scripts from the study
- the range of score points on the raw test score scale to cover
- the number of judges to involve
- the number of judgements to require each judge to make
- the number of packs of scripts to allocate to each judge
- the number of scripts from each test to include in each pack
- the range of test marks to be covered by the scripts in each pack
- the amount of overlap in the ranges of test marks covered by the scripts across packs
- whether to impose any extra constraints, for instance to minimise exposure of the same script to the same judge across packs; or to ensure that each script appears the same number of times across packs

The number of possible comparisons within a set of objects is given by (N objects*(N objects-1))/2. Given that the models used to process RO data allow for missing data, the RO judgment designs tended to be sparser and achieve between around 20% (e.g. Black and Gill, ibid.) and 50% (e.g. Black and Bramley, ibid,) of possible comparisons between the scripts available in a study. In different studies, this was achieved using different combinations of the above-mentioned features, in particular varying the number of packs of scripts and number of judges.

The studies carried out to date tended to include scripts on most of the mark range in a test while avoiding the scripts with the lowest number of marks, as this has been shown to increase the stability of the outcomes (see Bramley and Gill, 2010). Bramley and Gill also suggest that mark points likely to be in the vicinity of the grade boundaries of interest should be well within the mark ranges included in the study for the same reason.

Interestingly, recent research by Verhavert et al. (ibid.) failed to find an effect of the number of objects to be compared in CJ exercises on the reliability level reached. In other words, it appears that it is possible to reach reliable results with a CJ assessment regardless of the number of performances included in the assessment, as long as a sufficient number of comparisons per performance are made for the required level of reliability. In our context, this suggests that having scripts on fewer mark points from the effective mark range, perhaps 50% or fewer, as long as these were distributed sufficiently widely around the grade boundaries of interest, could be a way to optimise judging designs and make them more efficient.

However, ideally, the sample of scripts used in CJ exercises should be in some way representative of the full set of scripts from the relevant examination (cf. Benton, 2019). Reducing the number of mark points included in CJ exercises would potentially reduce the representativeness of the sample further, potentially leading to grade boundary outcomes that would not be representative of the outcomes that would have been obtained if the full set of scripts was judged. Furthermore, this could further exacerbate the interpretability of the bootstrapping results (see more on this later in this section). While the number of scripts to be included in CJ exercises may be limited by practical considerations in operational contexts, the precise impact of different sample profiles and sizes requires further research.

There has been some debate regarding the extent to which judges can more or less successfully judge scripts that display more or less 'balanced' performance (in the sense of low variability of individual unit total marks, or of section or question total marks within a paper) (Scharaschkin and Baird, 2000; Gray, 2000). Bramley (2012) found some evidence that perceptions of quality are affected by where and how the marks have been gained. The most notable effect on perceived quality came from the presence of blank (missing) responses. In this study, scripts with missing responses were perceived to be worse than those with incorrect responses and the same total score by the equivalent of 2 marks. Scripts with higher proportion of marks gained on questions testing 'good chemistry' or of marks gained on more difficult questions also tended to be perceived as better.

While this suggests that it would be best practice to try and select a 'representative' script or sample of scripts on each mark point for CJ exercises, defining what representative may mean in each case may be difficult, although the features considered by Bramley (ibid.) suggest a starting point for that. To our knowledge, no CJ studies explicitly selected scripts representative in the above sense, though some studies have used balanced scripts in terms of best fit to the Rasch model on the assumption that this would be helpful to judges and less likely to lead to biased results. For instance, Raikes, Scorey and Shiell (ibid.) used Rasch analysis to identify the scripts that best fitted the Rasch model in terms of their score profiles (removing those misfitting the model, i.e., containing a higher proportion of unexpectedly good answers to difficult questions or unexpectedly poor answers to easy questions).

Given that the main advantage of the RO method is in the efficiency with which a large number of paired comparisons can be derived from a relatively small number of judges and packs of scripts, the majority of the studies have used packs of 10 scripts, each pack deriving 45 paired comparisons[6] (e.g. Bramley, 2005; Gill, Bramley and Black, 2007; Black and Bramley, 2008). Other studies trialled smaller packs, with the rationale that they might be psychologically and physically easier for judges to rank order. Black and Gill (2008) used packs of 6 scripts, while Black (2008) and Raikes, Scorey and Shiell (ibid.) used 3 scripts per pack. The number of scripts per pack also determined the number of scripts from each session, with even numbers from each session in packs of ten or six scripts, and necessarily uneven number in packs of 3 scripts.

The studies have also varied the number of judges used. Bramley (2005) used 12 judges, Gill, Bramley and Black (ibid.) used 7, Black and Bramley (ibid.) used 9, Black and Gill (ibid.) used 3. In these studies, the judges tended to be senior

[6] (N scripts/N scripts – 1)/2.

examiners or experienced markers. Raikes, Scorey and Shiell (ibid.), however, used a more diverse range of judges, as well as larger groups of judges where each judge made only a small number of comparisons. They used the following groups of judges: members of the existing Awarding Committee (N=6), paid for 16 hours of work per person; other examiners that had marked the scripts operationally (N=48); teachers that had taught candidates for the examinations but not marked them (N=57); and university lecturers that teach the relevant subject to first year undergraduates (N=54). Each person in the last 3 groups was paid for 2 hours of work. They found very high levels of intra-group and inter-group reliability for the scales and measures estimated from all 4 groups' judgements as well as close agreement between grade boundary estimates, concluding that all groups made very similar judgments, and that members of all 4 groups could take part in a CJ exercise for setting grade boundaries without compromising reliability.

Their findings are confirmed by Verhavert et al. (ibid.), who found no effects of the number of judges, number of judgements per judge, or their expertise on how reliable CJ assessments can be. An effect was found only in the sense that 'novice' judges, i.e. judges with no content or marking expertise, needed a significantly higher number of comparisons per performance to achieve the same level of reliability as expert judges or peers familiar with subject content. In our context, this would suggest that expert markers (irrespective of amount of experience) as well as teachers and other content experts, could be expected to produce overall similarly reliable judgements.

Verhavert et al. (ibid.), however, make a point that including a large number of judges, while arguably leading to higher validity by way of including a wider range of opinions without harming reliability, could lead to a higher chance of potentially deviating judges being included. The extent to which deviating judges who misfit the Rasch model significantly affect the reliability is as yet unknown. It would also be interesting to explore to what extent misfitting judges would affect mark-measure correlations and grade boundary recommendations in the context of using CJ for standard maintaining.

Bramley and Gill (ibid.) evaluated how reducing the number of judges, the number of scripts per pack and the overlap between packs affected the outcome of a RO study. The outcomes investigated were separation and reliability; the correlation between test score and measure; and the effect on the substantive result (the equivalent test marks at the grade boundaries). Their findings, as well as the success of most other RO studies in deriving plausible script quality scales, suggest that the method is robust in that outcomes are fairly constant when factors such as the setting of the exercise, the number of judges, the number of scripts per pack are varied. They suggest that future RO exercises could possibly use fewer judges or fewer scripts per pack (around 25% fewer data points than were used in the original study analysed (7 judges and ten scripts per pack). However, Verhavert et al. (ibid.) demonstrate that the number of comparisons per object may be the main determinant of the outcome of a CJ exercise, and, while savings and optimisation may be possible in some respects, a sufficient number of comparisons per script for a desired level of reliability needs to be ensured in order to have confidence in the outcome.

Scale separation reliability

A key way of establishing whether a CJ exercise has worked is to check the properties of the scale of 'perceived quality' created by the judges. This involves investigating scale separation reliability (SSR) and model fit, which are the usual checks conducted for any latent trait analysis (cf. Bond and Fox, 2007).

The SSR coefficient is analogous to the 'person separation reliability' in Rasch modelling (e.g. Andrich, 1982) and to KR-20, Cronbach Alpha, and the Generalizability Coefficient. It is calculated as:

$$\text{SSR} = \frac{(\text{Observed SD})^2 - \text{MSE}}{(\text{Observed SD})^2}$$

where Observed SD is the standard deviation of the estimated measures, and MSE is the mean squared standard error of the estimated measures across all the scripts.[7]

In this context, SSR means 'reproducibility of relative measure location' (cf. Winsteps Manual https://www.winsteps.com/winman/reliability.htm; Verhavert et al., 2018). In our context, high reliability of the script measure scale would mean that there is a high probability that those scripts estimated with high measures actually do have higher measures than the scripts estimated with low measures.

In general, the decision of whether SSR of a scale can be considered satisfactory will depend on the purpose for which the scale is constructed, as well as on the context and type of the assessment under consideration. Verhavert et al. (2019) cite 0.7 as the level mentioned in the literature as appropriate for low-stakes or formative assessments, and 0.9 as the level often accepted as appropriate for high-stakes and summative assessments (Nunnally, 1978). In the RO studies carried out to date, SSR of around 0.8 and higher has generally been judged as satisfactory and related to other aspects of the CJ exercises being judged as satisfactory too.

Model fit

A common way of checking overall model fit is to check the overall proportion of misfitting judgements. Usually, this should be at or below what would be expected by chance, i.e. less than about 5% of standardised residuals using the criterion of 2 for the absolute value of the standardised residual, and less than about 1% using the criterion of 3 (cf. Linacre, 2011).

In addition to that, it is necessary to check the usual Rasch fit statistics for the scripts and judges (Wright and Linacre, 1994)[8]. In determining what might be considered appropriate infit and outfit mean square values for CJ data, it might be worth considering the range of 0.4-1.2 (i.e. a degree of overfit to the model) as appropriate

[7] Separation coefficient is the ratio of the person true SD (i.e., the 'true' standard deviation), to RMSE, the error standard deviation. It provides a ratio measure of separation in RMSE units, which is easier to interpret than the reliability correlation, with no upper bound as with SSR. Separation coefficient is the ratio of 'true' variance to error variance. The relationship between separation coefficient and SSR is: separation coefficient = square-root(SSR/(1-SSR)) (cf. https://www.winsteps.com/winman/reliability.htm).

[8] Note the limitations of Rasch-based fit statistics with respect to unknown exact sampling distributions (e.g., Christensen, et al., 2013; Karabatsos, 2000; Smith, Schumacker and Bush, 1998). However, useful applications of these indices have been demonstrated in the literature (e.g. Wright and Linacre, 1994), and their use for exploratory or descriptive purposes may be considered appropriate despite the limitations (e.g., Engelhard, Kobrin and Wind, 2014).

(cf. https://www.rasch.org/rmt/rmt83b.htm). This is because agreement levels might be increased compared to model expectation by virtue of the CJ method itself as well as by the judges' notions of script quality being more likely to be shared and constrained given their extensive familiarity with the relevant mark schemes.

While, according to Wright and Linacre (1994), somewhat lower fit statistics may be appropriate and expected in the context of judged data where agreement is encouraged, these can also be seen as less productive for measurement in that they indicate less information is gained from some observations and, in the case of low outfit mean squares, indicate imputed responses (Linacre, 2002). Furthermore, they may produce misleadingly high reliabilities and separations (Wright and Linacre, ibid.). While the extent of any overestimation due to overfit is unclear, this should be borne in mind when considering the SSR and separation cut off points as evaluation criteria for CJ methods. In their simulation study, Chambers, Vitello and Rodeiro (2019) found that CJ SSRs are overinflated compared to 'true' reliability (defined as squared correlation of CJ measures with simulated true marks) in exercises with around 10-15 judgements per script. While it is not clear what this overestimation is related to, nor that the two measures of reliability measure the same thing, these findings suggest caution when interpreting scale reliability as an indicator of success of smaller CJ exercises in particular.

Mark vs. measure correlation

Bramley and Gill (2010) demonstrated that the correlation between test score and measure that emerges from the latent trait analysis is not an artefact of the design and that random rankings lead to correlations close to zero. Thus checking the mark-measure correlation is a way to establish whether the judges in a CJ exercise were perceiving a trait that is sufficiently similar to the one underlying the test marks. Previous RO studies tended to consider correlations around and above 0.7 as satisfactory.

In subjects where marking reliability is generally high, we can be reasonably confident that a low mark-measure correlation indicates that the judges were perceiving a different trait to the one underlying the test marks (Bramley and Gill, 2010), which would cast doubt on the validity of the CJ measure scale. In subjects with known low marking reliability, however, low mark-measure correlation could suggest that original mark scale may have been at issue, or that the underlying trait is difficult to judge consistently, irrespective of whether marking or or comparative judgement is used.

Quantifying uncertainty in outcomes

Bramley and Gill (ibid.) demonstrate that the uncertainty in the outcome of a CJ study, i.e. the estimated cut marks in the new examination session, can be quantified by bootstrap resampling. The outcome depends on the regression equations summarising the relationship between test score and measure. In order to quantify the variability in the outcome, it is necessary to quantify the variability in the intercepts and slopes of these lines, which can be done using bootstrap resampling (Efron and Tibshirani, 1993; cited in Bramley and Gill, ibid.; Banjanovic and Osborne, 2016).

Bramley and Gill (ibid.) note that there is more sampling variability in the intercept and slope of the line when there is less of a linear relationship between test score and measure. In other words, the more the original score scale and the perceived

quality scale are at odds, the less we can be confident that the judges were perceiving the same qualities that led to the original marks.

In a bootstrap procedure, a large number (often 1000) of random samples of size N (where N is the number of scripts from a particular year) are drawn (with replacement) from each test session script set, and the slope and intercept parameters of the regression of test score on measure are estimated for each sample. Then, for any particular score point on one test, the distribution of the equivalent score point on the other test can be obtained.

Using this procedure, Bramley and Gill (ibid.) obtained uncertainty estimates from two previous RO exercises on KS3 English and AS psychology. The interquartile ranges (IQRs) obtained through bootstrapping showed that the middle 50% of marks corresponding to a given test score fell in a relatively narrow range – about 2 score points for the KS3 English, and about 3 score points for the AS psychology. They noted that the full range of the distributions could be quite wide, particularly at the extremes (over 21 score points in the case of AS psychology at the cut-score of 20). They also noted that there was more variability at the cut marks at the extremes of the score scale, and advise that the key cut marks to be estimated from CJ studies should ideally not be at the extremes of the score scale.

It should be noted that larger samples used in bootstrapping tend to produce better precision and narrower confidence intervals (Banjanovic and Osborne, ibid.). Typical sample sizes in the context of CJ exercises would be about 50 (scripts/mark points). This could lead to wide confidence intervals which may underestimate the actual precision of the CJ outcome.

Bramley and Gill also point out that any observed variability in the outcomes of bootstrapping as applied to CJ data should be considered and interpreted within the context of a particular exercise being evaluated. Estimated variability of a particular outcome does not answer questions about what might have happened with a different experimental design. The bootstrapping procedure treats the pairs of values (test score, measure) for each script as random samples from a population and shows what other regression lines might have been possible with other random samples from the same population.

Assumptions of CJ methods

An assumption of CJ and other judgemental methods when used for standard maintaining is that judges can allow for differences in test difficulty when judging script quality or that judges are able to compare between better performance on easier questions vs. worse performance on more difficult questions and make adjustments in their quality judgements accordingly. Beyond evaluating the outcomes of CJ exercises in ways described above, it is difficult to establish at this point whether this assumption is entirely justified.

Good and Cresswell (ibid.) questioned the ability of judges to make sufficient adjustments for differences in paper difficulty between two tiers. On the other hand, Black's (2008) RO exercise of scripts from different tiers suggests that the judges were able to do this with a reasonable level of agreement, as evidenced by SSRs of above 0.75 across four exercises. This suggests that the judges may indeed be able to make adjustments for differences in test difficulty to some extent. It should be noted, however, that in typical standard maintaining situations the judges are not normally required to make adjustments for large test difficulty differences of the

magnitude that are present between tiers. Furthermore, as Benton and Bramley (ibid.) point out, between-tier equating (a kind of 'vertical equating') is problematic for any method, including statistical equating, and is not straightforwardly done or interpretable. The differences between exams in consecutive exam sessions should be fairly small as the test papers are normally based on quite tight test specifications, and thus presumably more straightforward to adjust for when making judgements about script quality.

Furthermore, previous research has shown that, while judges may not be able to precisely estimate empirical difficulty of individual test items (Impara and Plake, 1988; Pollitt et al., 2007), they seem able to judge relative difficulty of test items reasonably well. For instance, Impara and Plake (1988) found correlation of 0.78 and Holmes et al. (2018) of 0.76 between judged difficulty and empirical difficulty (see also, e.g. Brandon, 2014; Attali, Saldivia, Jackson, Schuppan, and Wanamaker, 2014; Curcin, Black and Bramley, 2010; etc. for further evidence of this). Arguably, in the context of CJ exercises, an ability to see differences in item difficulty in relative terms may be sufficient for valid outcomes that sufficiently capture the effects of test difficulty differences on test performance.

Another relevant assumption of CJ methods relates to the statistical models which are used to analyse the data, which require that each paired comparison is independent of the others (assumption of local independence). This assumption is not met by design in the RO method given that the paired comparisons are extrapolated from, and thus constrained by, the rank order in each pack. While it is possible that reliability indices may be somewhat inflated due to violations of local independence in each method and especially in RO, as Bramley (2007: 279) points out, it may be that if the objects to be ranked are sufficiently far apart on the psychological scale then many of the possible outcomes of comparisons derived from a rank order in particular would be so unlikely as to have effectively a zero probability, making the constraint imposed by a ranking in practice much less than it seems in theory. This problem might be further alleviated in the RO method by using the RO model rather than the paired comparisons Rasch model to analyse the RO data. It should be noted, however, that the PCJ method may also not be immune to violations of local independence, which are likely to be more pronounced than in psychological experiments where comparisons of simple traits such as perceived brightness are made. This is because the complex nature of the objects being compared, i.e. scripts, makes it unlikely that individual comparisons would be entirely independent of each other due to memory effects for individual scripts (Bramley, 2005: 204). This aspect of the problem may be alleviated by constraining the number of times each judge sees each individual script in a PCJ or a RO exercise.

Method

Overall study design and specifications used

The CJ pilots were conducted in 4 subjects which have been awarded for at least 3 years, namely, GCSE media studies, GCSE English language, AS English literature and AS psychology. Nine different specifications across the 4 subjects were included in the pilots. Table 2 shows the specifications used and the pilots that were conducted in each (more detail about each method piloted is provided later in this section). Except with media studies, where a specification from only one board was included, specifications from at least 2 exam boards were included for other subjects, for example, English literature 1 came from one board, and English literature 2 came from a different board.

The pilots were conducted at paper level for each specification. For each pilot, the aim was to equate by expert judgment the tests and performances from the previous examination session (year 1) and current examination session (year 2), focusing on the relevant key grade boundaries A and E in the AS specifications and A/7, C/4 and F/1 in GCSE specifications. Where possible, the pilots were conducted during live examination sessions and prior to awarding, while some were conducted outside of live marking and awarding.

Paper based RO and online PCJ[9] pilots with expert markers as judges were carried out for most specifications. A separate 'teacher' PCJ pilot was also carried out on one specification, using a wide pool of AS and A level teachers as judges.

Table 2 *Specifications and CJ methods used*

Specification	RO	PCJ	Pinpoint PCJ	Teacher PCJ
Media	✓			
Eng lit 1	✓			
Eng lit 2	✓		✓	✓
Psych 1	✓	✓	✓	
Psych 2	✓	✓		
Eng lang 1		✓		
Eng lang 2	✓			
Eng lang 3		✓		
Eng lang 4	✓	✓		

Some pilots were designed to test the lower bound of reliability that could be obtained from fairly small-scale exercises, which might be considered necessary if these methods were to be implemented operationally. Other pilots sought to collect more data, i.e. more comparisons per script, to establish whether this could lead to more consistently higher reliability. This was tried with GCSE English language, a qualification which consists of primarily extended response questions and consequently has lower marker reliability (Rhead, Black and Pinot de Moira, 2018) (and presumably performance qualities that may be more difficult to pin down even

[9] PCJ pilots were carried out online using the beta version of the NoMoreMarking™ website (www.nomoremarking.com).

when judged holistically). In this way, we sought to establish whether conducting CJ exercises on a larger scale in such a subject could lead to high reliability irrespective of possible inherent difficulty of judging the qualities of performances for individuals.

Participants

Table 3 below details the number of judges that took part in each pilot and exercise. In specifications where both RO and PCJ were conducted, the judges that did the RO also did the PCJ alongside a few other judges. In English language 4, all of the judges that did RO also did PCJ.

The expert judges were initially contacted through examination boards and invited to take part. Recruitment was on a 'first come, first served' basis, and the judges were required to have at least 2 years examining experience in at least one paper of the relevant specification as well as some teaching experience. A number of participants were senior examiners – team leaders, principal examiners, and a few were chairs of examiners. They were paid for their participation.

The teachers were mostly recruited via emails to schools' administration offices, who were asked to forward the invitation to the relevant department/teacher. The emails were sent to schools with ten or more AS candidates taking the relevant specification. The teachers were recruited on a first come, first served basis, and were required to have a minimum of 2 years teaching experience of the relevant qualification at AS or A level.

Table 3 *Number of judges per method and specification.*

Specification	RO	PCJ	Teacher PCJ	Pinpoint PCJ
Media	6			
Eng lit 1	6			
Eng lit 2	6		41	10
Psych 1	6	10		10
Psych 2	6	10		
Eng lang 1		15		
Eng lang 2	15			
Eng lang 3		20		
Eng lang 4	15	15		

Scripts

All the scripts used in the pilots were obtained from examination boards. The scripts were provided without candidate identifiers and all marks and annotations were removed.

The scripts used in PCJ pilots were the same as those used in the RO pilots, where both were conducted. For each paper within a specification, the judging designs[10] for these methods aimed to include one script per mark point from around 50-70% of the effective mark range, while excluding scripts with the lowest marks (first 10-14 marks for AS level, 5-6 for GCSE) as these were deemed more difficult to judge.

[10] See the section on judging designs below for more detail.

In the 'pinpoint' PCJ, however, which was trialled on English literature 2 and psychology 1, the judges only looked at scripts on or near the key grade boundaries (A and E).

For Media studies, English literature 1 and all English language specifications, where pilots were conducted either during the marking period or immediately after awarding, a random selection of scripts per mark point (one per mark point) was provided by the boards. For the specifications where pilots took place outside of the live marking and awarding period (English literature 2, psychology 1 and psychology 2), given that there was more time available to prepare the pilots, we first analysed the item level data from each examination session using the partial credit Rasch model to obtain student (script) fit statistics. We then randomly sampled one script per mark point (15 scripts per relevant mark point for the pinpointing PCJ) from the scripts with infit and outfit mean square statistics greater than 0.5 and less than 1.5. In this way, we hoped to obtain scripts with a good fit to the Rasch model in that they did not contain a high proportion of poor responses to easy questions and vice versa. This is one possible definition of 'balanced' scripts (cf. Raikes, Scorey and Shiell, ibid.) and it was hoped that this would have eased the judging process somewhat.

Table 4 shows the number of scripts per paper that were used in each pilot. Where there is only one row per specification in the table, the maximum number of marks, and hence number of scripts selected for the pilots, were the same for both papers within that specification.

Table 4 *The number of scripts per paper in CJ pilots*.

		N scripts	
Specification	Paper max mark	RO/PCJ/Teacher PCJ	Pinpoint PCJ
Media	80	56	
Eng lit 1	72	51	
	44	31	
Eng lit 2	50	35	30
Psych 1	72	53	30
Psych 2	75	51	
Eng lang 1	80	56	
Eng lang 2	80	56	
Eng lang 3	64	45	
	96	67	
Eng lang 4	80	40	

Once received, the scripts were checked for general legibility, labelled, and either printed and arranged into packs for RO pilots, or electronically uploaded to the No More Marking software for judging, as per the judging designs described below.

Judging designs

Table 5 shows the key features of the judging designs for each pilot at paper level. Where the paper is not identified by paper 1 (P1) or paper 2 (P2), the features were

the same across both papers. Where there was a difference between papers (Specification 2 and Specification 8), this was because the 2 papers in these specifications had different maximum marks.

For the RO pilots, given the relative efficiency of the method in deriving a large number of paired comparisons from a small number of rank orders, we aimed to achieve 20-35 comparisons per script. This derived 25-35% of the possible comparisons per paper and session (cf. Raikes, Scorey and Shiell, ibid.). In designing the RO judging allocation, care was taken to minimise same script exposure to judges, so a judge saw a particular script a maximum of 2 times. In addition, mark ranges of the scripts appearing in the same pack were constrained, with the average of 19-20 marks (min 5, max 29) in English language pilots, and with the average of 12-13 marks depending on paper (min 6, max 20 marks) in other pilots. The number of marks overlap between packs was usually 8-12 marks. In Media studies and English literature 1 pilots, the packs were presented to judges from worst to best (i.e. the earlier packs contained scripts with lower marks). In the other pilots, the packs were presented from best to worst (cf. figure 1). Each judge had a unique set of packs. An example of a pack design is presented in Appendix 1.

The PCJ pilots (including pinpointing), carried out on English literature 2, psychology 1 and psychology 2, given the amount of time required for judging and a relatively small number of judges that were made available to us, we achieved 10-12 comparisons per script. Despite knowing from other research that this would likely produce relatively low SSR levels, it was deemed useful to trial pilots on this scale to establish the lower bound of reliability achievable with quantities of data that might be inevitable in some operational contexts. In English language pilots, as mentioned previously, we trialled on the basis of a minimum of 20 comparisons per script, to establish the maximum reliability that might be possible in such a subject on the scale that might be pushing the boundaries of what might be possible operationally, but might still be achievable in certain contexts.

In all PCJ pilots except for English language, judges were allocated pairs of scripts randomly across the whole mark range available but the pairs were constrained in that they always contained one script from the previous session and one script from the current session. For English language, the pairings were completely random and comparisons within and across sessions were allowed.

For the pinpointing CJ exercise, the pairs were randomly created within the constraint that each pair contained scripts from different sessions, but were allocated separately for each grade boundary (separate ‘tasks’ on No More Marking platform were created for each grade boundary). Each pair of scripts allocated to judges contained one script on the previous session grade boundary and one script from a narrow range of mark points around the potential new session grade boundary. The range of mark points to include for the new session was determined based on the estimates from a prior ‘mini’ RO exercise. For this, 6 judges actually conducted a full RO exercise with 70% mark points included (reported in the results section later). From this, we selected 50% of the mark points and removed the surplus scripts and all comparisons involving them. We then analysed this smaller data set to determine preliminary estimates of the new session A and E grade boundaries in the same way as for the full data set. Three scripts on these preliminary boundaries, as well as 3 scripts per 2 mark points either side of them were included in the pinpoint PCJ (a total of 15 scripts per new session grade boundary). Fifteen scripts on each of the

previous session grade boundaries were also included. Therefore, in the pinpointing exercise, for each grade boundary, each pair of scripts seen by the judges contained one script on the previous session grade boundary (randomly selected with replacement from the set of 15 grade boundary scripts) and one script from the current session, randomly selected from the 15 available scripts selected as described above. A total of 180 judgements per grade boundary were collected, with 12 judgements per script.

Table 5 *Key features of judging designs by paper*

Method	Specification	N judgms per script	% mark range	N judges	N packs/ pairs per judge	Total N judgms (% poss comps)	N contracted days per judge
RO	Media	32 (20-35)	70	6	20	1800 (28%)	1.5
	Eng lit 1 P1	28 (20-30)	70		16	1440 (28%)	1
	Eng lit 1 P2	22 (20-25)	70		8	720 (38%)	0.5
	Eng lit 2	21 (20-25)	70		8	720 (30%)	0.5
	Psych 1	31 (30-35)	70		18	1620 (29%)	1.25
	Psych 2	30 (25-35)	70		17	1530 (30%)	1.25
	Eng lang 2	36 (35-40)	70	15	9	1960 (32%)	0.6
	Eng lang 4	22 (20-25)	50	15	4	800 (25%)	0.25
PCJ	Psych 1	10	70	10	53	530 (10%)	0.8
	Psych 2	10	70		51	510 (10%)	0.8
	Eng lang 1	20	50	15	75	1120 (18%)	1.3
	Eng lang 3 P1	25	70	20	59	1170 (29%)	1
	Eng lang 3 P2	25	70	20	87	1742 (19%)	1.5
	Eng lang 4	20	50	15	69	1040 (33%)	1.2
Pin PCJ	Eng lit 2/Psych 1	12	N/A	10	18 (per boundary)	180 (10%)	0.5
Teacher PCJ	Eng lit 2	12	70	41	10	410 (17%)	0.25

Procedure

For each pilot, the judges were provided with detailed instructions, clean copies of examination papers and mark schemes, electronic recording forms to record their judgments of paper difficulty, electronic recording forms to record their ranks (for RO pilots), and a survey to complete after completing their tasks. Sample instructions for RO and PCJ are included in Appendix 2. In RO pilots, these materials, alongside the packs of scripts, were sent to judges by post. In PCJ pilots, all the documents and the links to the judging software were sent by email.

The judges were asked to (re-)familiarise themselves with the relevant examination papers before starting each task. They were also asked to form a judgment about the relative difficulty of the papers from different examination sessions and record this on the form provided. They were asked to keep this difference in mind and take it into account when judging the quality of student responses.

The judges were instructed not to use the mark schemes provided to remark the papers. It was emphasised that these were provided for reference (for example, in case they were not sure about the correct response to a question).

In the RO pilots, the judges were instructed to place the scripts in each pack into a single rank order, from best (rank 1) to worst (rank 6). In the PCJ pilots, they were supposed to decide which of the pair of scripts were better. In both cases, the judges were asked to judge holistically, in terms of overall quality, while taking into account differences in difficulty between the exam papers. For each exercise, the judges were given 5 to 6 days to complete them and return the results electronically.

Data analysis

The data from RO pilots were analysed using the FACETS software version 3.66.3 (Linacre 2010). For the PCJ data, initial analysis was provided automatically by the No More Marking platform. The data was additionally analysed using the Sirt package in R (Robitzsch, 2019) which implements the Bradley-Terry model (Bradley and Terry, 1952), to obtain more detailed statistics and where it was deemed necessary to exclude some misfitting judges and/or judgements before rerunning the analysis.

For all pilots except for the pinpointing, the judgemental grade boundary estimates were then obtained using the linear regression procedure described earlier. For the pinpoint PCJ, the grade boundaries were obtained using the logistic regression (cf. Benton and Elliott, 2016). In this case, grade boundary estimates were derived directly from paired comparisons (i.e. the measure quality scale, although estimated for judgement evaluation purposes, was not used for estimating grade boundaries). The grade boundary mark was estimated to be the mark on the Y2 scripts where the modelled probability of the Y2 scripts beating the Y1 grade boundary scripts was greater than or equal to 0.5.

Standard rounding rules were applied throughout. In each case where pilots were carried out for both papers in a specification, paper-level grade boundaries were first estimated, and then combined to obtain qualification boundaries. It should be noted that in the current awarding procedures SRBs are statistically estimated at specification level, and then paper-level boundaries are estimated based on this information alongside the input from awarders' judgements.

Some data cleaning was undertaken in the following ways:

- most misfitting observations were removed from analyses (based on highest standardised residuals, usually those higher than absolute 4). In order to preserve most of the data, we tended to remove misfitting observations rather than all judgements from a judge that showed some misfit
- in a small number of cases, the all of the judgements of one or two judges were removed
- all scripts which won or lost all their comparisons, and hence had imputed measures, were removed from regression and mark-measure correlation analyses
- in a few cases, particularly outlying scripts were also removed from regression and mark-measure correlation analyses

The results were evaluated in terms of model fit, scale properties, mark-measure correlation, plausibility of grade outcomes, and uncertainty in grade boundary estimates as described previously.

An analysis was also conducted to compare judgement consistency when judges make within session comparison vs. between session comparisons, as it could be hypothesised that misfit might be higher in the latter case as these would be the more difficult judgements to make, threatening the validity of the resulting script quality scale. Following the procedure described in Pollitt (2015), average weighted mean squares (infit mean squares) for between and within session judgements were calculated for each data set containing paired comparisons both within and between sessions (English language 1, 3 and 4). Large discrepancies between these would suggest possible issues with interpreting the overall scale of script quality, especially if the between-session comparisons were found to lead to significantly less consistent judgements compared to those within session.

Some additional analyses were also carried out to investigate whether further optimisation of the judging designs might be possible, for instance reducing the profile or the number of judges or the number of scripts (mark range) in the pilots (cf. Bramley and Gill, ibid.). Thematic qualitative analysis was also undertaken to analyse some aspects of judges' survey responses.

Results

In this section we first present overall evaluation of the results across all the specifications and pilots, and then go on to present the grade boundary estimates and evaluation of those in terms bootstrap confidence intervals for individual specifications.

Evaluation of Rasch model fit and scale properties

In order to have confidence in the results of a RO exercise, it is necessary to evaluate model fit and script quality scale properties in particular (see Data analysis section for more details on this).

Model fit

Overall model fit can be seen in Table 6 for each specification, paper and pilot. In general, model fit was satisfactory, with around 5% or less standardised residuals greater than absolute 2 and around 1% or less standardised residuals greater than absolute 3.[11]

As can be seen from figures 3 and 4 below, individual judge fit was satisfactory in general, with most infit and outfit mean squares between 0.4 and 1.4, suggesting that the judges were reasonably consistent in their judgements. There were a few judges with fit statistics significantly higher than average even after some of their most misfitting judgements were removed, suggesting unpredictable or unexpected judging behaviour compared to other judges. We have tried rerunning the relevant analyses after excluding these judges entirely and found that this had virtually no impact on the outcomes in terms of grade boundary estimates in these cases. Therefore, these judges were retained for the final analyses.

In addition, there were a number of judges with low infit and oufit mean square statistics of under 0.4, who overfit the Rasch model, i.e. exhibit more predictable judgement patterns than is expected by the model. This was particularly prominent in PCJ data. While the scale separation and SSR tended to be higher in RO exercises than in PCJ, the average infit and outfit mean squares tended to be closer to model expectation of 1 in RO exercises, whereas in the PCJ exercises these were lower (except in the case of pinpointing PCJ, where the patterns were more similar to those in RO). A similar pattern is also apparent in script fit statistics.

It is unclear why the two methods should exhibit these different patterns with respect to fit. Speculating, these patterns may be to some extent related to the way scripts were allocated to judges. In PCJ, the scripts were randomly paired in terms of their original score, which resulted in pairs where there were large differences between original scores, and hence overall quality too. It would have been more likely for judges to overwhelmingly or completely agree about the winners and losers in such pairs, which would have led to low fit statistics in those cases, reducing the average fit. On the other hand, in the RO exercises the packs of scripts are put together in such a way that the range of marks in each pack is reasonably small, possibly leading to lower likelihood of judges always agreeing on certain comparisons. The small mark ranges included in the pinpointing PCJ exercises could similarly explain

[11] Appendix 3 contains the details of data cleaning undertaken to improve model fit in some cases. Judge fit statistics and script statistics are presented in appendices 4 and 5 respectively.

the higher fit statistics there. Both of these would lead to less consistency and higher fit statistics. Furthermore, in RO, due to the way pairs are extrapolated from ranks, a misaligned rank order within a pack could affect a lot of derived paired comparisons, leading to higher fit values, while the local independence violations would still lead to stretching of the scales and overestimation of separation and SSR for that reason to some extent.

Table 6 *Overall model fit.*

				StRes > abs 2		StRes > abs 3	
Specification	Paper	Exercise	N valid observations	N	%	N	%
Media	P1	RO	3584	166	4.63	50	1.40
Eng lit 1	P1	RO	2870	128	4.50	26	0.90
	P2		1440	62	4.31	18	1.25
Eng lit 2	P1	RO	1432	56	4.18	26	1.67
	P2		1434	60	4.31	12	0.83
	P1	Teacher PCJ	854	14	1.64	2	0.23
	P2		806	32	3.97	2	0.25
	P1-A	Pin PCJ	360	14	3.89	0	0.00
	P1-E		360	10	2.78	2	0.56
	P2-A		360	14	3.89	4	1.11
	P2-E		360	8	2.22	2	0.56
Psych 1	P1	RO	3240	152	4.69	46	1.42
	P2		3214	132	4.11	38	1.18
	P1	PCJ	1060	12	1.13	0	0.00
	P2		1060	16	1.51	0	0.00
	P1-A	Pin PCJ	360	10	2.78	0	0.00
	P1-E		360	20	5.56	2	0.56
	P2-A		360	6	1.67	0	0.00
	P2-E		360	16	4.44	0	0.00
Psych 2	P1	RO	3046	120	3.94	30	0.98
	P2		3056	122	3.99	44	1.44
	P1	PCJ	1020	12	1.18	2	0.20
	P2		1020	20	1.96	4	0.39
Eng lang 1	P1	PCJ	2246	10	0.45	4	0.18
	P2		2246	20	0.89	8	0.36
Eng lang 2	P1	RO	4016	152	3.78	28	0.70
	P2		4032	110	2.73	34	0.84
Eng lang 3	P1	PCJ	2356	32	1.36	8	0.34
	P2		3470	58	1.67	14	0.40
Eng lang 4	P1	RO	1778	42	2.36	14	0.79
	P2		1786	46	2.58	12	0.67
	P1	PCJ	2064	30	1.45	12	0.58
	P2		2060	36	1.75	10	0.49

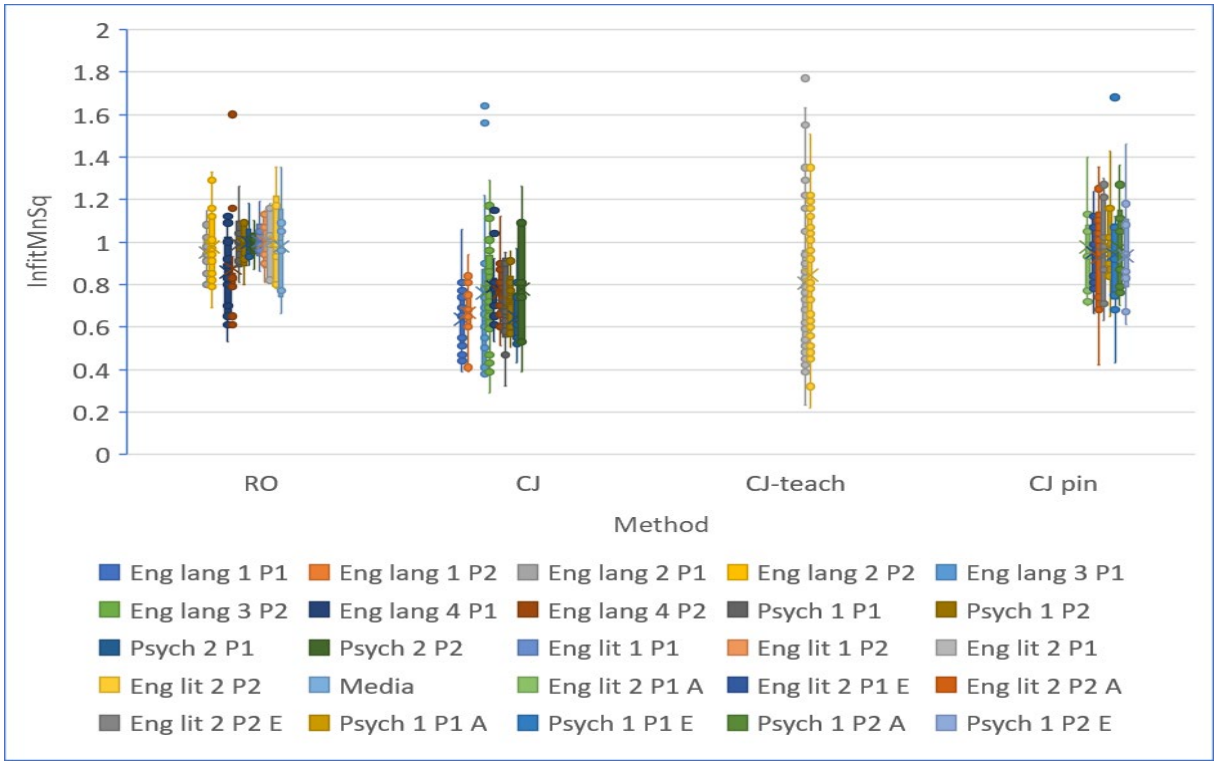


Figure 3 *Judge infit by paper and method*

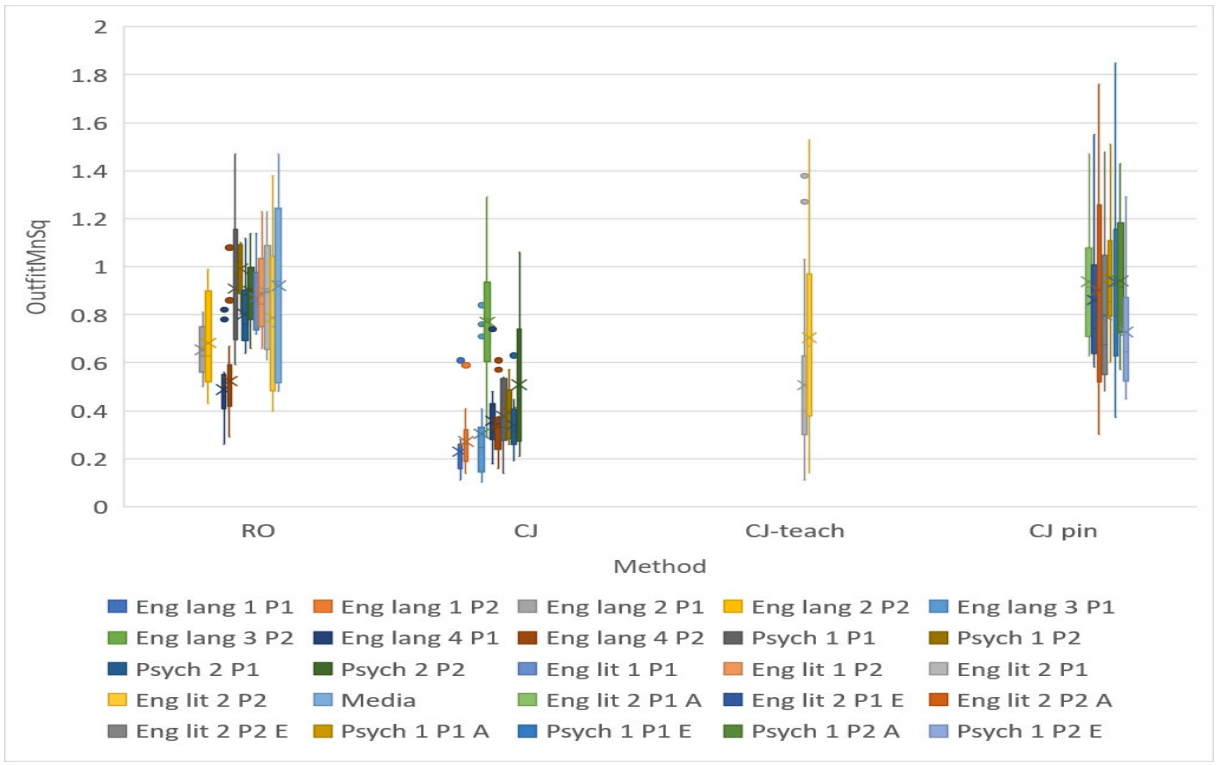


Figure 4 *Judge outfit by paper and method*

As previously mentioned, low fit statistics may be a sign that the SSRs and separations are to some extent overestimated although high standard errors that are assigned to imputed measures, or measures arising from near perfect agreement, should to some extent limit the amount of overestimation of separation and SSR both in RO and PCJ exercises. While the extent of reliability overestimation is unclear, it is

unlikely that this would be over and above the scale of differences in reliability between the PCJ pilots based on 10 comparisons per script (SSRs 0.7-0.8) vs. those with 20 or more comparisons per script (SSRs > 0.9) – see table 7 below. Indeed, Chambers, Vitello and Rodeiro (ibid.) found that the extent of overestimation compared to 'true' reliability may be greater for CJ exercises with fewer judgements per script. This suggests that we can be reasonably sure that even if there was some level of overestimation in our larger-scale pilots, the reliability would be unlikely to drop below acceptable evaluation threshold of around 0.8. However, it may be possible that some of the reliability and separation coefficients from the smaller pilots may be overinflated to the extent that they may in some cases have dropped below our minimum SSR thresholds for acceptability had they not be overestimated.

The partitioning of misfit for the 3 pilots where this was possible indicates that there were no major discrepancies in consistency for within- and between-session judgements. The figure below shows the result of this analysis for each paper of the 3 specifications where both within and between-session comparisons were available. The analysis was carried out and is presented for both original and cleaned data in each case.

It can be seen that although there was somewhat more inconsistency in the between-session judgements, the differences tended to be very small. Even in the case of EL4 paper 1, where the differences were slightly larger, the infit mean square for the between-session comparisons was well below the model expectation of 1, suggesting that the judgements were still sufficiently consistent and unlikely to have degraded the measurement scale.

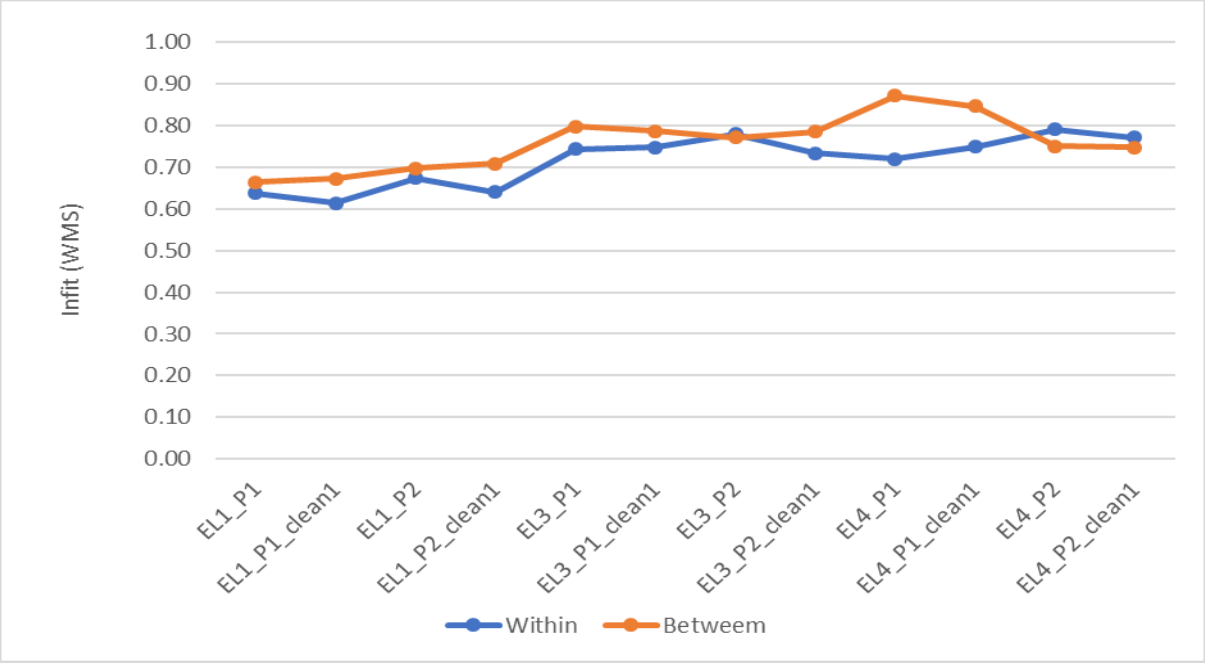


Figure 5 *Partitioning of infit mean squares for within- and between-session judgements*

Scale properties

The SSR and separation as well as mark-measure correlations for each specification, paper and pilot are presented in Table 7 below. We discuss the

pinpointing results separately at the end of this section, and they are not included in the graphs below and the accompanying discussion.

Table 7 *SSR and separation coefficients*

Specification	Paper	Exercise	SSR	Separation	Mark-measure correlation Y1	Mark-measure correlation Y2
Media	P1	RO	0.98	6.00	0.91	0.92
Eng lit 1	P1	RO	0.95	4.38	0.79	0.83
	P2		0.92	3.33	0.75	0.68
Eng lit 2	P1	RO	0.89	2.88	0.55	0.59
	P2		0.92	3.51	0.79	0.72
	P1	Teacher PCJ	0.76	1.94	0.68	0.71
	P2		0.69	1.79	0.60	0.71
	P1-A	Pin PCJ	0.35	1.24		0.11
	P1-E		0.61	1.60		0.20
	P2-A		0.64	1.66		-0.19
	P2-E		0.67	1.74		0.05
Psych 1	P1	RO	0.98	6.85	0.95	0.94
	P2		0.94	4.03	0.92	0.90
	P1	PCJ	0.79	2.19	0.87	0.84
	P2		0.75	2.10	0.78	0.84
	P1-A	Pin PCJ	0.35	1.24		-0.01
	P1-E		0.61	1.59		0.23
	P2-A		0.34	1.23		0.23
	P2-E		0.73	1.92		0.25
Psych 2	P1	RO	0.96	5.01	0.90	0.90
	P2		0.97	5.87	0.93	0.96
	P1	PCJ	0.79	2.21	0.70	0.89
	P2		0.76	2.04	0.89	0.80
Eng lang 1	P1	PCJ	0.93	3.73	0.91	0.95
	P2		0.93	3.73	0.91	0.93
Eng lang 2	P1	RO	0.99	8.97	0.93	0.91
	P2		0.98	6.48	0.94	0.90
Eng lang 3	P1	PCJ	0.95	4.51	0.93	0.95
	P2		0.95	4.34	0.92	0.94
Eng lang 4	P1	RO	0.97	5.94	0.90	0.92
	P2		0.97	6.06	0.94	0.95
	P1	PCJ	0.95	4.69	0.96	0.95
	P2		0.94	4.20	0.94	0.95

It can be seen that, with the exception of the pinpointing pilots, in most other cases the SSRs were close to 0.8 or above. This would suggest that most pilots with reasonable number of comparisons per script and a relatively wide mark range of scripts included in them succeeded in producing reproducible measures of quality. A

further evidence of this comes from the relationship of measures produced in pilots carried out for the same specifications (see Table 8), which were highly correlated, even in cases where the number of comparisons per script was around 10. Where different CJ methods were used for the same specifications, the scripts used were the same, but allocated to judges in different combinations between methods where the judges were the same between methods. In some cases not all or none of the judges were the same.

Table 8 *Correlations between measures of quality from parallel pilots*

Specification	Pilot 1	Pilot 2	Same judges?	Measure correlation P1	Measure correlation P2
Eng lit 2	RO	Teacher PCJ	no	0.70	0.60
Psych 1	RO	PCJ	6/10	0.89	0.88
Psych 2	RO	PCJ	6/10	0.90	0.90
Eng lang 4	RO	PCJ	yes	0.93	0.95

There was more variability in mark-measure correlation compared to the SSRs, with correlations of 0.55-0.60 in a few other pilots. However, the majority of the mark-measure correlations where above 0.7.

As expected based on previous research, there appears to be a strong relationship between the number of comparisons per script and the SSR and separation (correlation with average number of comparisons of 0.86 across pilots for both SSR and separation).

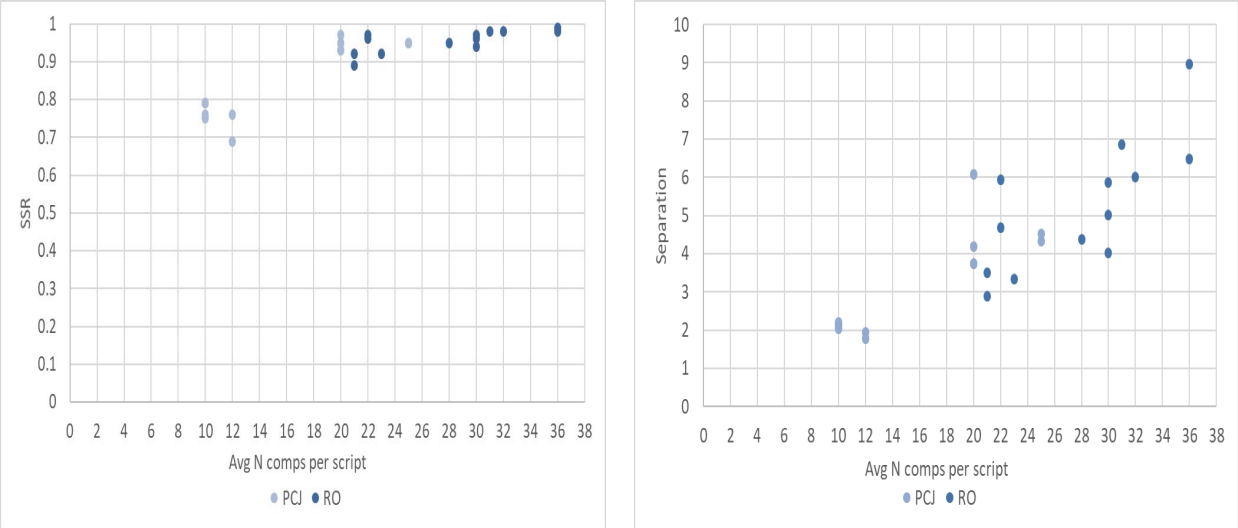


Figure 6 *SSR and separation by average number of comparisons per script*

However, there does not seem to be such a strong relationship between mark-measure correlation and number of comparisons. Figure 7 plots mark-measure correlations by average number of comparisons and method. The banners for each dot represent the relevant SSR. Across methods, the correlation between the number of comparisons and mark-measure correlation was 0.44. This was higher for correlation between mark-measure correlation and SSR/separation at 0.60.

As can be seen from the figure below, at around 10 comparisons per script, and SSR at 0.7-0.8 (the SSRs are denoted by the banners next to each data point), the mark-measure correlations range from around 0.7 to 0.9. At 20 comparisons or

more, the correlations are in most cases above 0.9. However, as denoted in the figure, this was not the case with English literature, where, irrespective of the method, and at similar number of comparisons per script and SSR, the mark-measure correlations do not reach the correlations that are achieved in the other subjects.

This pattern with English literature could point to either some issues with marking reliability in the AS English literature papers, or with some issues with the mark scheme, and how well the performance qualities rewarded by the mark scheme aligned with the performance qualities considered by the judges in our pilots. This kind of problem could not be entirely overcome by increasing the number of comparisons per script in the pilots, and should probably not be considered to be an issue with judging reliability in the pilots, which was reasonable in terms of SSR and comparable to other subjects at similar average number of comparisons per script.

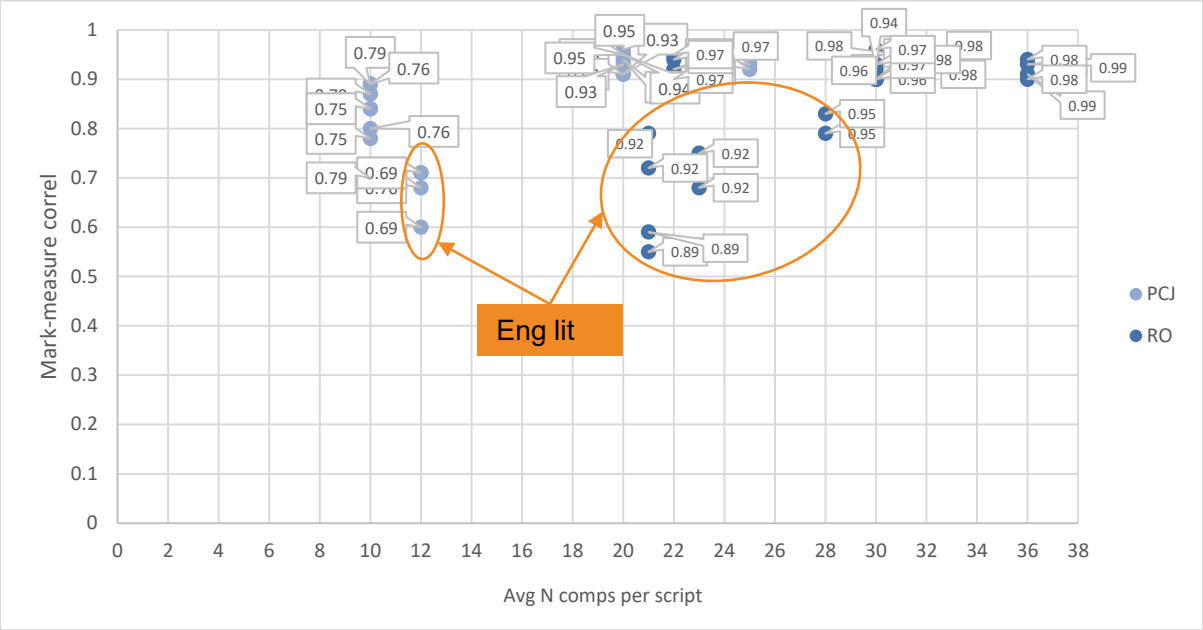


Figure 7 *Mark-measure correlations by average number of comparisons per script*

It should be noted that using the paired-comparisons model to analyse the rank-ordering data can lead to some over-estimation of the statistical separation of the objects on the latent trait because the rankings constrain the possible paired comparisons outcomes, leading to violation of the assumption of local independence in the model (cf. Bramley, 2005). There is currently no reliable estimate of the amount of over-estimation this causes. However, the English lang 4 pilots, where the same judges compared the same scripts (in different combinations) via both the rank-ordering and the PCJ method, and the average number of comparisons per script was similar in each method, suggest that the level of overestimation may not be so large as to be of concern, with separation difference between methods of 1.2 to 1.8. Furthermore, both RO and paired comparisons may violate local independence in terms of probably unavoidable effects of memory of different scripts, that could affect outcomes of both rank orders and paired comparisons to some extent. One way to overcome statistical overestimation might be to use the RO model to analyse RO data.

In pinpointing pilots, the reliability and separation were less than what might ideally be expected, particularly for grade A. In this case in particular, we cannot be

confident that the observed differences between scripts were not due to measurement error. Furthermore, agreement between the mark and measure scale in these pilots was very low. This may be to some extent unsurprising given the small range of mark points, and thus, small differences in script quality, of the scripts that the judges saw at each grade boundary (cf. Bramley, 2007). However, this also casts additional doubt on the ability of the judges to create a meaningful script quality scale within such a small range of marks, at least in an exercise where each script was seen a relatively small number of times (N=12).

Grade boundary results

The sections below are organised by specification to aid specification-level comparisons between methods. Where more than one pilot was conducted for the same specification, the results are presented for each pilot separately, followed by the summary tables of grade boundary estimates.

As described in more detail in the data analysis section, the grade boundaries for the current session were produced using linear regression of mark on measure obtained from the pilots, except in the pinpointing exercise, where current grade boundaries were obtained using logistic regression. In each case, the paper level grade boundaries for the current session (Y2) were estimated first. These were then combined, applying any relevant weighting factors, to estimate the qualification level boundaries. It should be noted that the paper level grade boundaries are operationally set only after the qualification level boundaries had been set, which may account to some of the differences between pilot and operational boundaries presented below.

In a few cases, some scripts ‘won’ or ‘lost’ all of their comparisons and thus had imputed measures with large standard errors. These scripts did not contribute to measure estimation of other scripts. Where there were two or more such scripts, it was deemed appropriate to remove them from the regression analysis as their imputed measures could not be deemed to represent a realistic measure of quality. Similarly, a few outlier scripts were removed from the regression in a few specifications.

Using the bootstrap resampling procedure described previously, we attempted to quantify uncertainty in the outcome of the pilot. As a way of getting a sense of the variability in the estimated grade boundary outcomes, we present the ranges based on middle 50% range of possible results, as well as +/- 2SD around the mean estimate. It is unclear at the moment which range might be more appropriate for this context, though we would suggest that middle 50% might be sufficient given prior research showing that these methods are reasonably stable when altering different features incidental to the design itself.

For instance, Black and Bramley (2008) found that the outcome of a study was replicated when the exercise was carried out postally (i.e. with judges working alone at home) compared to when all the judges were together in a face-to-face meeting. Bramley and Gill (2010) showed that manipulating features such as number of judges or size of packs in a RO study had relatively little effect on final grade boundary outcomes until the data sets are eroded too much. In the current research, in almost all cases where more than one pilot was conducted on the same specifications (sometimes with the same judges, sometimes with different judges),

the results largely replicated each other closely, with high correlations between resulting quality measures. Furthermore, our data manipulation analysis presented later in this report also suggests that the results are reasonably robust based on data from different groups of judges, and different number of comparisons per script.

Therefore, it could be argued that the standard +/- 2SD range may be too pessimistic as an indicator of the replicability of the outcomes, especially within the constraints of exercises which may be carried out year on year with similar design parameters and judge profiles. We suggest that middle 50% inter quartile range (IQR) might be appropriate, and we base our discussion on that, though we present the wider +/- 2SD ranges throughout for reference.

In the current pilots, the judges were asked to form a judgment about the relative difficulty of the papers from different examination sessions (Y1 more difficult, Y2 more difficult, similar) and record this on the form provided. They were asked to keep this difference in mind and take it into account when judging the quality of student responses. We present the frequency of responses of each judge group alongside grade boundary results in each section and consider whether their initial judgements of paper difficulty differences between sessions agree with the outcomes of the CJ exercises in terms of implied paper difficulty differences. It should be noted that these normally include very small number of judgements of empirical test difficulty that is notoriously difficult to judge reliably even at item level, let alone at test level. Therefore we would not expect a high level of agreement either between judges or with the grade boundary outcomes.

Media studies

Table 9 presents the regression equations for calculation of marks, including grade boundary marks, in Y2 corresponding to equivalent performance in Y1. The mark-measure relationship is also presented in Figure 8, alongside the distribution of equivalent marks in Y2 corresponding to each of Y1 judgemental grade boundaries obtained through bootstrapping.

Table 9 *Regression equations for calculation of Y2 grade boundaries – Media studies*

Y	Equation	R2
Y1 mark	45.45+3.8*x	0.83
Y2 mark	50.23+3.47*x	0.84

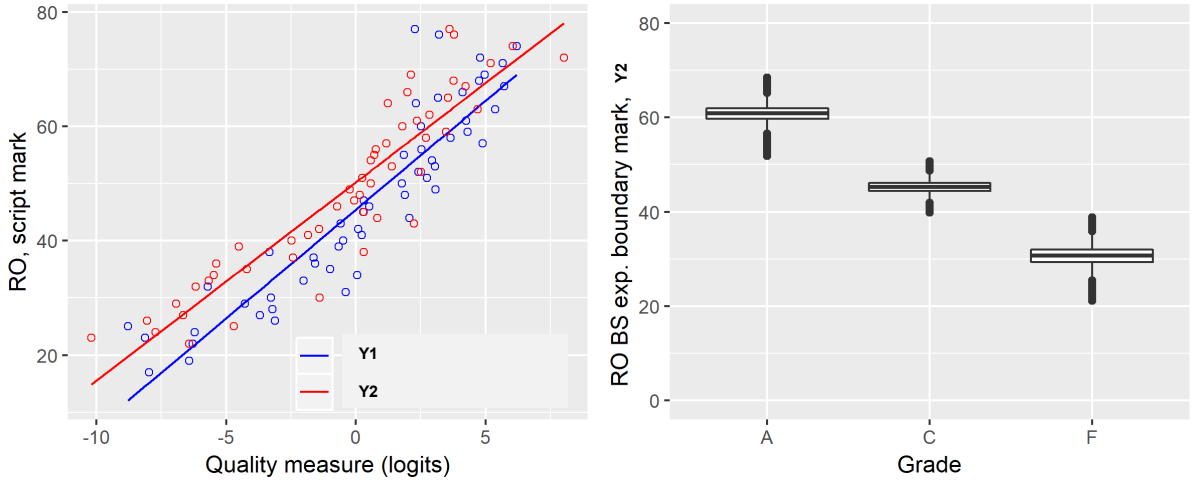


Figure 8 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – Media studies*

The interquartile ranges show that the middle 50% of marks corresponding to grade A and C boundaries score fell in a narrow range of around 2 marks. This was 3 marks for grade E. This additionally suggests that the results obtained in the RO pilot are credible.

Table 10 shows operational and RO pilot grade boundaries, the latter calculated from the above-mentioned equations. It can be seen that the pilot grade boundaries are very close to the operational ones in Y2. They are similarly higher than the Y1 boundaries, suggesting that to show similar performance quality the students had to score higher marks. This suggests that the Y2 test was easier.

The RO grade boundaries were within the tick chart ranges for all grade boundaries. The bootstrapping exercise also suggests some potential variability in the RO boundaries, with the middle 50% ranges of 2-3 marks depending on grade boundary.

Table 10 *Operational and pilot judgemental grade boundaries – Media studies*

Boundaries	A	C	F
Y1 operational	57	40	24
Y2 operational	62	46	31
Y2 RO pilot	61	45	31
Y2 t/c	60-65	44-48	28-32
50% IQR	60-62	44-46	29-32
2SD	3	3	4

Table 11 shows how the judges involved in this exercise saw paper difficulty differences between sessions. Recall that they were asked to make this initial judgement about paper difficulty differences before starting on the CJ tasks and were asked to account for these differences in their quality judgements. All the judges thought that the Y2 paper was more difficult, which agrees with the results of the RO pilot.

Table 11 *Judges' initial views of paper difficulty differences – Media studies*

	Y1 more difficult	Y2 more difficult	Papers similar	Total
P1	6	0	0	6

Taken together, the results of this pilot suggest that the RO exercise succeeded in producing credible grade boundaries, based on a plausible script quality scale and high level of agreement between test score and script quality measure scale.

English literature 1

Table 12 presents the regression equations for calculation of marks, including grade boundary marks, in Y2 corresponding to equivalent performance in Y1. The mark-measure relationship is also presented in Figures 9 and 10, the distribution of equivalent marks in Y2 corresponding to each of Y1 judgemental grade boundaries obtained through bootstrappig.

Table 12 *Regression equations for calculation of Y2 grade boundaries – English literature 1*

	Y	Equation	R2
P1	Y1 mark	41.21+4.6*x	0.62
	Y2 mark	39.68+5.17*x	0.68
P2	Y1 mark	26.52+3.07*x	0.47
	Y2 mark	25.72+2.87*x	0.57

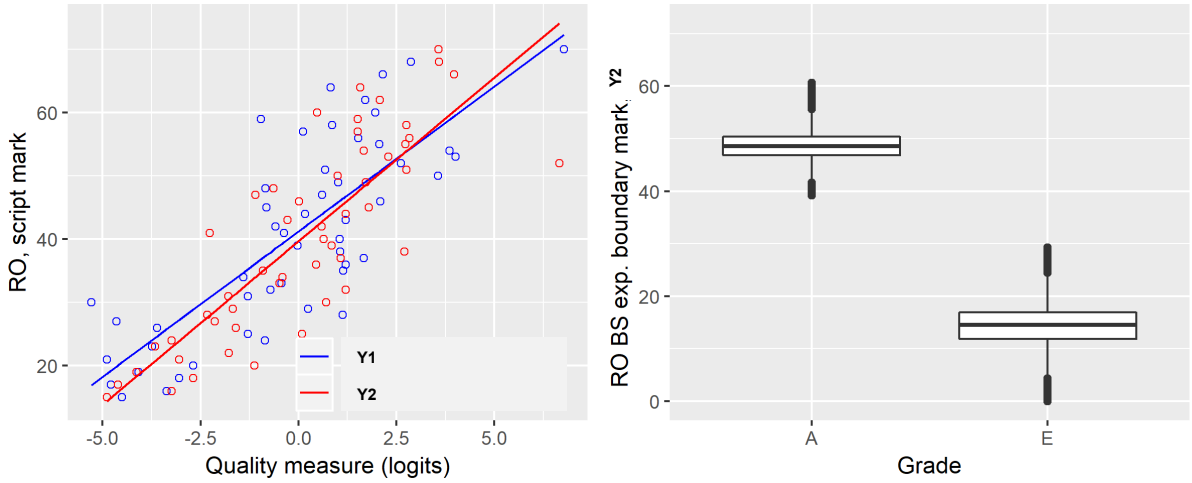


Figure 9 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – English literature 1 P1*

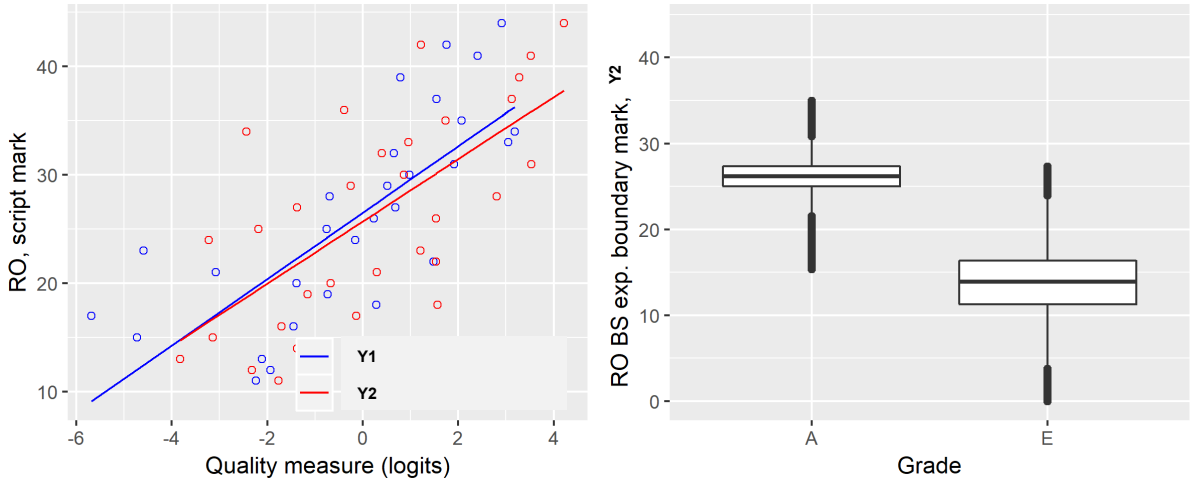


Figure 10 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – English literature 1 P2*

For paper 1, the interquartile ranges show that the middle 50% of marks corresponding to grade A boundary score fell in a narrow range of 3 marks and 5 marks for grade E. For Paper 2, this was 2 marks for grade A and 5 marks for grade E.

Table 13 shows that the pilot grade boundaries are within 2 score points of the operational Y2 ones for A boundary at paper level, and identical at qualification level. For E grade, the pilot boundaries were up to 3 score points away from operational Y2 boundaries at paper and qualification level.

The RO grade boundaries were within the tick chart ranges for all grade boundaries. The bootstrapping exercise also suggests some potential variability in the RO

boundaries, with the middle 50% ranges of 2-3 marks for grade A, but around 5 marks for E grade in each paper.

Table 13 *Operational and pilot judgemental grade boundaries – English literature 1*

	P1		P2		Overall[12]	
Boundaries	A	E	A	E	A	E
Y1 operational	49	19	27	14	78	34
Y2 operational	46	18	28	14	77	33
Y2 RO pilot	48	15	26	14	77	30
Y2 t/c	43-49	15-21	25-31	11-17		
50% IQR	47-50	12-17	25-27	12-17		
2SD	5	7	3	7		

Table 14 shows that the views of paper difficulty differences were quite mixed and not clearly related to the outcome of the CJ exercise in terms of paper difficulty differences, particularly for paper 2. For paper 1, the majority of judges thought that the papers were similar between sessions, which does suggest the grade A RO outcome, but is less reflective of the grade E outcome.

Table 14 *Judges' initial views of paper difficulty differences – English literature 1*

	Y1 more difficult	Y2 more difficult	Papers similar	Total
P1	1	0	5	6
P2	3	0	3	6

Taken together, the results of this pilot suggest that the RO exercise produced credible grade boundaries, based on a plausible script quality scale and a reasonably high level of agreement between test score and quality measure scale, particularly for paper 1. Based on our evaluation criteria, we could be more confident in the results for paper 1, even though the departure from the operational Y2 grade boundaries is larger for this paper.

The results for paper 2, however, give some reason to be more cautious about interpreting that particular result, especially given slightly lower mark-measure correlations and resulting larger variability in possible grade boundary estimates. This is despite the apparent higher agreement between operational and pilot grade boundaries in this paper compared to paper 1.

English literature 2

RO and teacher PCJ

Table 15 presents the regression equations for calculation of grade boundary marks in Y2 corresponding to equivalent performance in Y1 for the RO and PCJ exercises respectively. The mark-measure relationship is also presented in Figures 11 to 14, alongside the distribution of equivalent marks in Y2 corresponding to each of Y1 grade boundaries obtained through bootstrapping.

[12] Weighting factor for paper 2 is 1.091.

Table 15 *Regression equations for calculation of Y2 grade boundaries – English literature 2*

		Y	Equation	R2
RO	P1	Y1 mark	31.75+2.70*x	0.30
		Y2 mark	33.75+2.68*x	0.35
	P2	Y1 mark	31.80+3.52*x	0.63
		Y2 mark	33.83+2.75*x	0.52
PCJ	P1	Y1 mark	32.36+5.08*x	0.46
		Y2 mark	33.20+4.49*x	0.50
	P2	Y1 mark	32.55+4.47*x	0.36
		Y2 mark	34.37+4.69*x	0.51

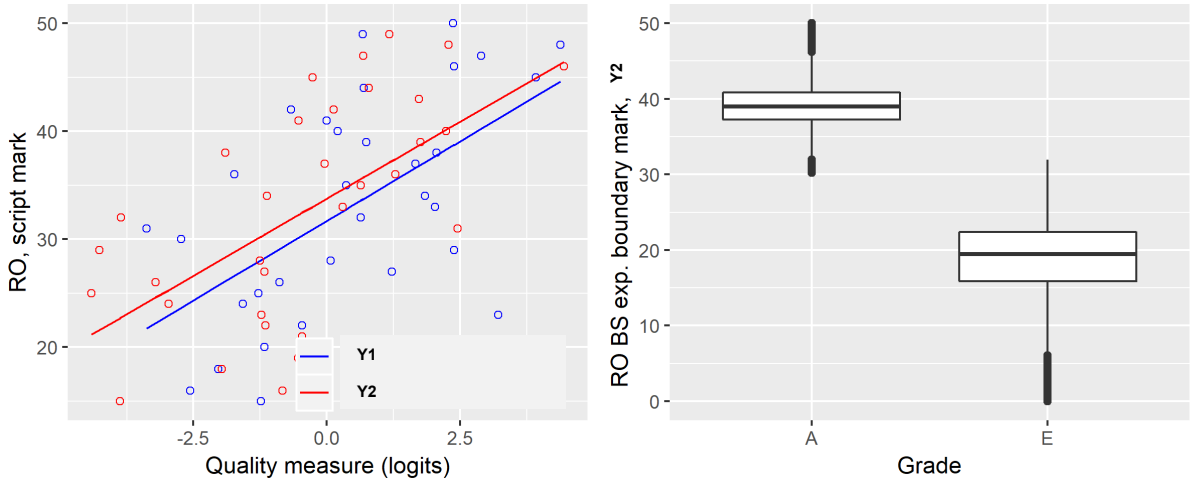


Figure 11 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – English literature 2 P1 – RO*

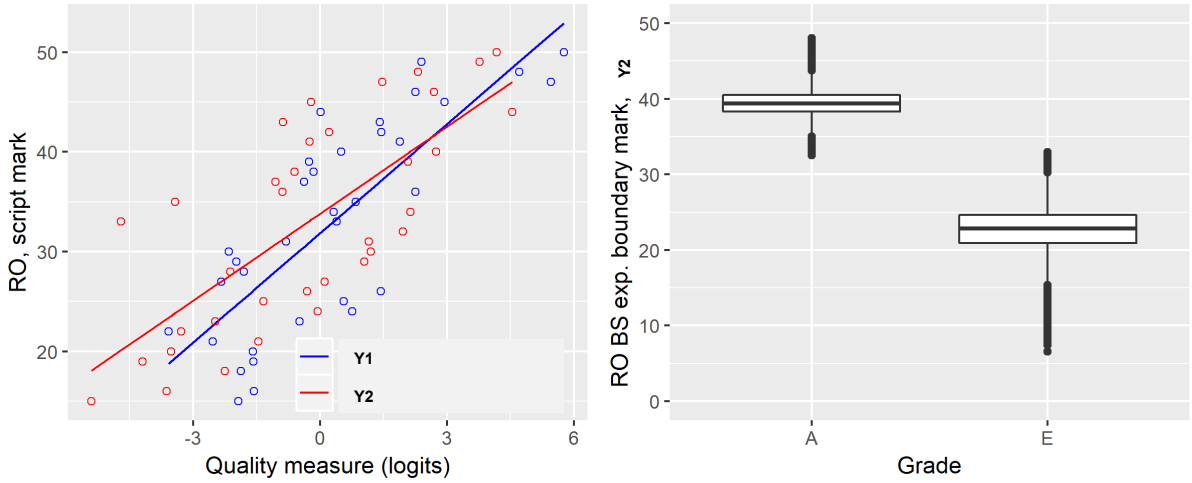


Figure 12 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – English literature 2 P2 – RO*

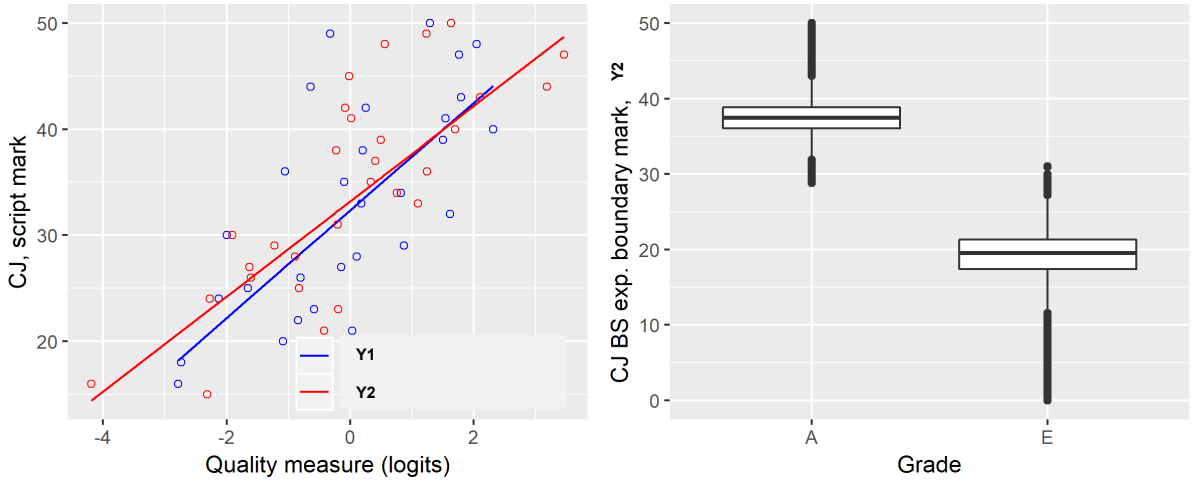


Figure 13 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – English literature 2 P1 – teacher PCJ*

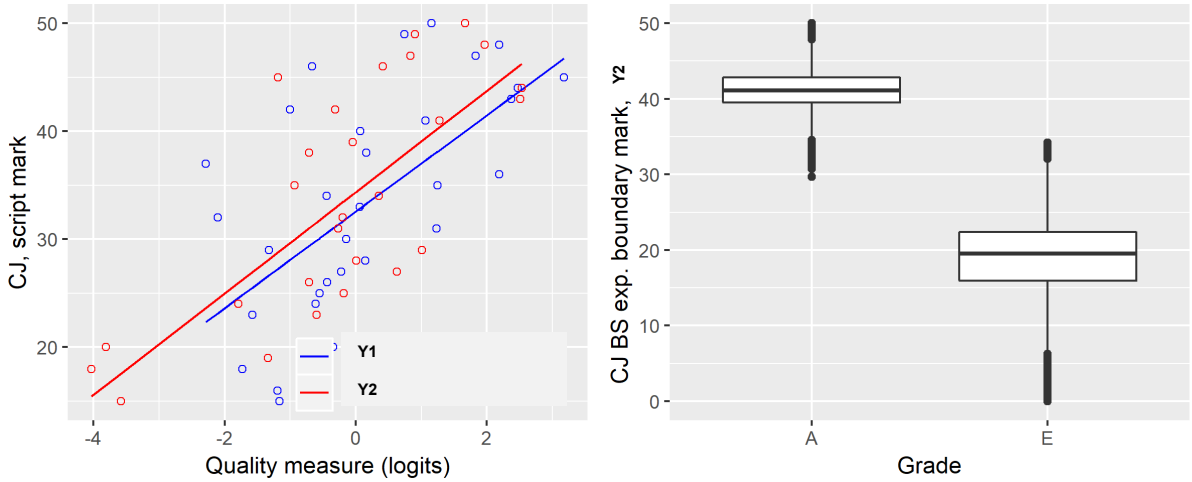


Figure 14 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – English literature 2 P2 – teacher PCJ*

For the RO exercise, for paper 1, the interquartile ranges show that the middle 50% of marks corresponding to grade A boundary score fell in a narrow range of 3 marks. This was 7 marks for grade E. For Paper 2, the middle 50% range of marks for grade A was 2, and 7 for grade E.

For the PCJ exercise, for paper 1, the interquartile ranges show that the middle 50% of marks corresponding to grade A boundary score fell in a narrow range of 3 marks. This was 4 marks for grade E. For Paper 2, the middle 50% range of marks for grade A was 3, and 7 for grade E.

Table 16 shows that the pilot grade boundaries are within up to one mark of the operational Y2 ones for A boundary at paper level and qualification level. For E grade, the pilot boundaries were 3 score points away from operational paper 1 Y2 boundary and 7 score points away from the paper 2 operational Y2 boundary. At qualification level, the E boundary is 10 score points away from the operational boundary.

The teacher pilot A grade boundaries for both papers are very close to both operational and RO pilot A grade boundaries. For E grade on paper 1, the teacher and RO pilot boundaries agreed exactly, but for the paper 2 E grade boundary, the teacher pilot result was in between the operational and RO pilot result. At qualification level, the A grade boundaries agree almost exactly, whereas for E grade, the two pilot boundaries are closer to each other than to the operational boundary. Furthermore, the measures from RO and PCJ pilots were reasonably correlated at 0.70 for P1 and for 0.60 P2, suggesting that both exercises produced fairly similar script quality scales in terms of script rank order, although less well correlated than those where both were produced by largely the same participants (see psychology 1 and psychology 2 below).

For RO pilot, the A grade estimates were within the tick chart ranges, while the E grade estimates were outside. For the teacher PCJ pilot, only the paper 1 A grade estimate was within the tick chart ranges. There was more overlap when comparing the tick chart ranges with our 50% IQR ranges. It should be noted that the IQR ranges were fairly wide for E grade boundaries at 4-7 marks depending on paper and exercise. This further suggests that we could not have the same amount of confidence in E grade boundary results as we could in those for A boundary.

Table 16 *Operational and pilot judgemental grade boundaries – English literature 2 – RO*

	P1		P2		Overall	
Boundaries	A	E	A	E	A	E
Y1 operational	37	17	39	18	76	35
Y2 operational	39	16	38	16	77	32
Y2 RO	39	20	39	23	78	42
Y2 PCJ	37	20	41	19	78	39
Y2 t/c	37-41	14-18	36-40	14-18		
50% IQR RO	37-41	16-23	38-40	21-25		
50% IQR PCJ	36-39	17-21	39-43	15-22		
2SD RO	6	11	3	5		
2SD PCJ	4	7	5	10		

Table 17 shows that the judges tended to see the papers from the 2 sessions as similar. In addition, about a quarter of the teachers thought that Y1 papers were more difficult. This is related to some of the RO and CJ outcomes but not all (e.g., P1 grade E boundary would suggest that the Y1 paper was more difficult, and grade A boundary that the papers were reasonably similar).

Table 17 *Judges' initial views of paper difficulty differences – English literature 2*

Pilot	Paper	Y1 more difficult	Y2 more difficult	Papers similar	Total
RO	P1	1	1	4	6
	P2	1	0	5	6
PCJ	P1	13	7	20	40
	P2	12	8	20	40

In the RO pilot, similarly to English literature 1, there was a relatively prominent difference in the overall quality and credibility of the outcomes by paper. Here, the scale reliability as well as mark-measure correlations were substantially higher for paper 2, with the variability in outcomes based on bootstrapping hence less pronounced for paper 2, particularly for grade A. Yet, while the grade A pilot

boundary was very close to the operational Y2 boundary for paper 1, the grade E boundary is quite discrepant. Indeed, the operational boundary is outside the middle 50% of the marks produced by bootstrapping.

In the case of paper 1, while the reliability of the scale might still be considered reasonable, the mark-measure correlations are probably lower than would be ideal, and lead to quite a high level of variability in potential outcomes, based on bootstrapping. Thus, even though both A and E pilot grade boundaries are within reasonable distance from the operational Y2 boundaries, grade E in particular may be suspect, given the issues with mark-measure correlation. Again, more evidence would be needed before a decision regarding the appropriate E grade boundary could be reached.

Similarly, in teacher PCJ pilot, the scale reliability as well as mark-measure correlations were higher for paper 1, with the variability in outcomes based on bootstrapping more pronounced for paper 2, particularly for grade E. Both grade A and E operational values were not within the 50% interquartile range for paper 2 and the teacher pilot boundaries are discrepant with the RO pilot boundaries. In this case, given better reliability and correlations for paper 2 in the RO exercise, the grade boundary results for paper 2 from this exercise might be considered more credible.

Pinpointing PCJ

As described previously, the starting point for the 'pinpoint' PCJ was a 'mini' RO exercise based on 50% of the mark range, which provided initial grade boundary estimates. Three scripts on these preliminary boundary marks, as well as 3 scripts per 2 mark points either side of them were included in the pinpoint PCJ (a total of 15 scripts per Y2 grade boundary). Fifteen scripts on each of the Y1 grade boundaries were also included. The pinpointing exercise was carried out separately for each paper and key grade boundary. According to initial mini RO estimates, Y2 grade boundaries were as follows:

Table 18 *Operational and mini RO judgemental grade boundaries – English literature 2 – pinpointing*

	P1		P2	
Boundaries	A	E	A	E
Y1 operational	37	17	39	18
Y2 operational	39	16	38	16
Y2 mini RO	39	19	41	20

Therefore, the following mark points were included in the pinpointing pilot:

- for paper 1 grade A: 37, 38, 39, 40, 41
- for paper 1 grade E: 17, 18, 19, 20, 21
- for paper 2 grade A: 39, 40, 41, 42, 43
- for paper 2 grade E: 18, 19, 20, 21, 22

Because the logistic regression analysis, which was used to estimate the Y2 grade boundaries in this pilot does not rely on Rasch measures specifically but rather on the probability of whether a new session script beat the previous session script, we did not exclude those scripts that won or lost their comparisons from the regression analysis. The grade boundaries for Y2 were calculated from the equations displayed in the table below as described in the data analysis section.

Table 19 *Regression equations for calculation of Y2 grade boundaries – English literature 2 – pinpointing*

Boundary	Equation
P1_A	0.52 + -0.01*x
P1_E	-3.06 + 0.13*x
P2_A	3.72 + -0.09*x
P2_E	-0.44 + 0.02*x

Table 20 *Operational and pinpoint PCJ judgemental grade boundaries – English literature 2 – pinpointing*

	P1		P2		Overall[13]	
Boundary	A	E	A	E	A	E
Y1 operational	37	17	39	18	76	35
Y2 operational	39	16	38	16	77	32
Y2 RO pilot	39	20	39	23	78	42[14]
Y2 teacher PCJ	37	20	41	19	78	39
Y2 pinpoint PCJ	47	23	42	20	89	43

Given the small sample of scripts used for each grade boundary, we did not consider it appropriate to evaluate variability in grade boundary estimates using bootstrapping. However, taking into account the reliability levels achieved for each grade boundary, and the comparison of the grade boundary outcomes of pinpointing and the other 2 pilots, shown in the table above, it can be seen that paper 1 grade A estimate based on pinpointing is unlikely to be correct. Paper 1 grade E and paper 2 grades A and E are more aligned with the outcomes of the other 2 pilots, and based on more reasonable scale reliabilities, so could be considered more trustworthy.

Psychology 1

RO and PCJ

Table 21 presents the regression equations for calculation of marks, including grade boundary marks, in Y2 corresponding to equivalent performance in Y1. The mark-measure relationship is also presented in Figures 15 and 18, alongside the distribution of equivalent marks in Y2 corresponding to each of Y1 grade boundaries obtained through bootstrapping.

Table 21 *Regression equations for calculation of Y2 grade boundaries – psychology 1*

		Y	Equation	R2
RO	P1	Y1 mark	44.15+4.49*x	0.90
		Y2 mark	42.71+4.20*x	0.89
	P2	Y1 mark	41.44+5.56*x	0.84
		Y2 mark	40.72+4.63*x	0.81
PCJ	P1	Y1 mark	44.10+6.51*x	0.75
		Y2 mark	42.05+5.89*x	0.70
	P2	Y1 mark	39.37+5.92*x	0.61
		Y2 mark	40.77+6.40*x	0.70

[13] Weighting factor for paper 2 is 1.091.
[14] This value is due to rounding.

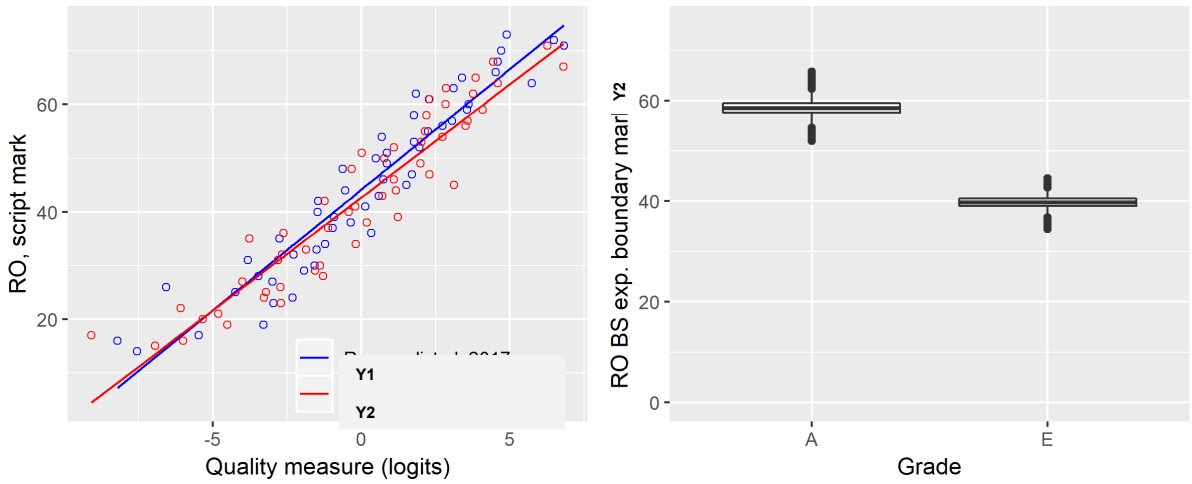


Figure 15 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – psychology 1 P1 – RO*

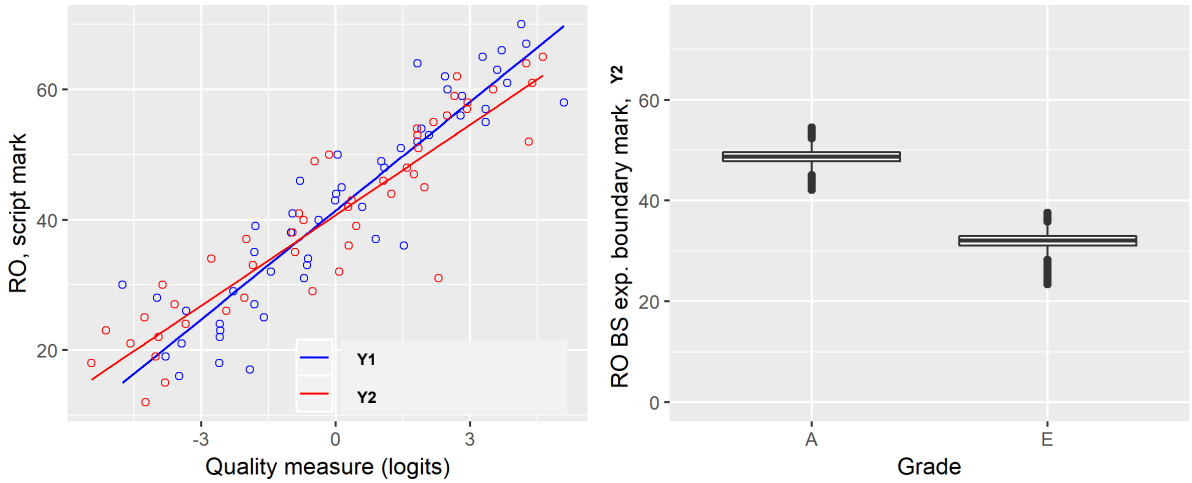


Figure 16 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – psychology 1 P2 – RO*

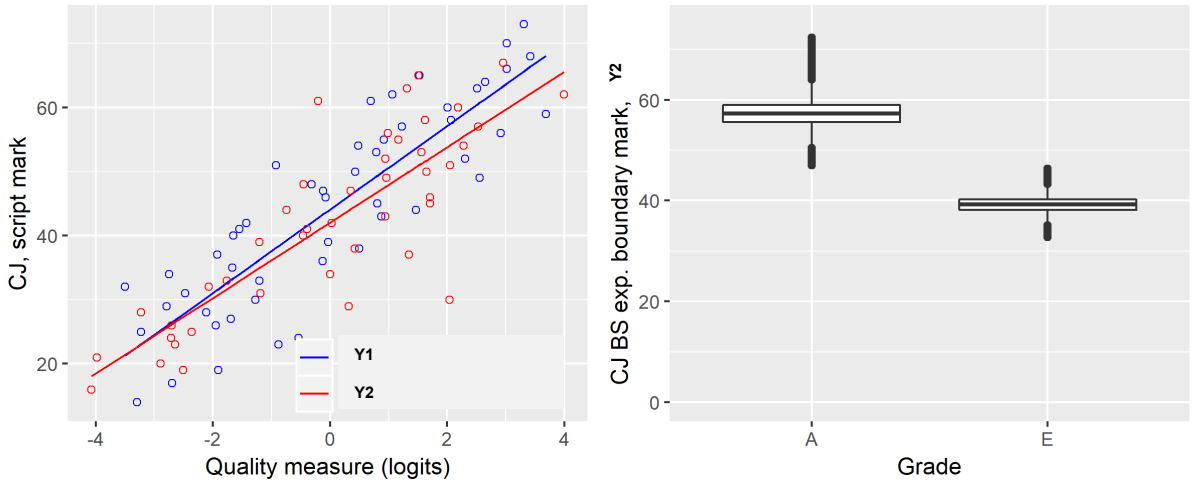


Figure 17 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – psychology 1 P1 – PCJ*

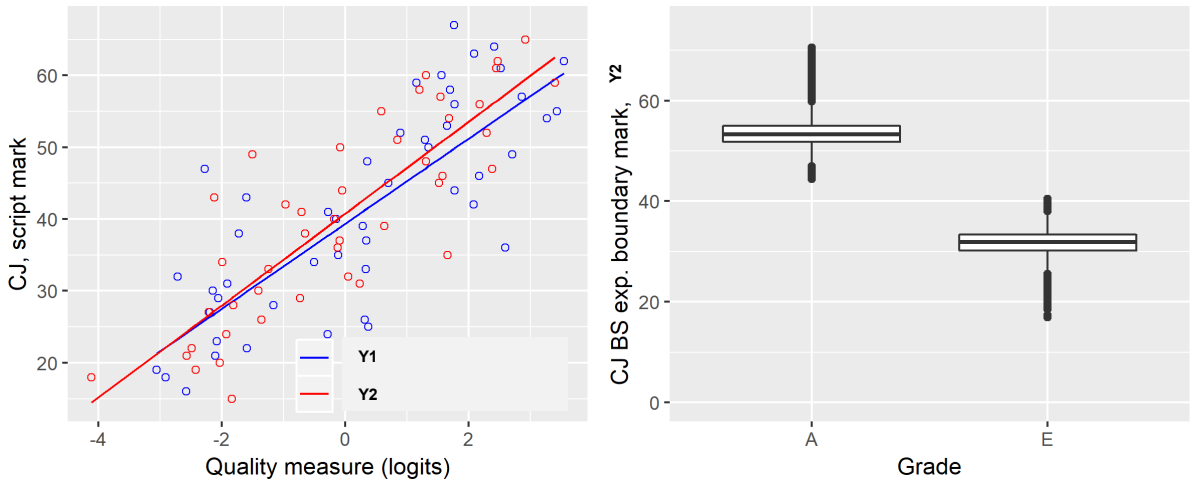


Figure 18 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – psychology 1 P2 – PCJ*

For RO, for paper 1, the interquartile ranges show that the middle 50% of marks corresponding to grade A boundary score fell in a narrow range of 2 marks. This was 1 mark for grade E. For paper 2, the middle 50% range of marks for each grade was 2.

For PCJ, for paper 1, the interquartile ranges show that the middle 50% of marks corresponding to grade A boundary score fell in a narrow range of 3 marks. This was 2 marks for grade E. For Paper 2, the middle 50% range of marks was 3 for each grade.

Table 22 shows that the RO pilot grade boundaries are within 2-3 marks of the Y2 operational ones for A boundary at paper level and 5 marks at qualification level. For E grade, the pilot boundaries were 8 marks away from operational paper 1 Y2 boundary and 4 marks away from the paper 2 operational Y2 boundary. At qualification level, the E boundary is 8 marks away from the operational boundary.

The PCJ pilot grade boundaries for both papers are close to the RO pilot boundaries except for 4-mark difference at paper 2 grade A. The PCJ and RO pilot boundaries both differed from the operational Y2 boundaries to the similar extent and in the same direction, further suggesting that the operational boundaries may have been too low. Furthermore, the measures from RO and PCJ pilots were highly correlated at 0.89 for P1 and for 0.88 P2, suggesting that both exercises produced similar script quality scales in terms of script rank order.

Unsurprisingly, the pilot boundaries were outside the tick chart ranges in all cases except for paper 1 grade A. The bootstrapping exercise also suggests some potential variability in the pilot boundaries, with the middle 50% ranges of 1-3 marks depending on grade boundary and pilot. It can be seen that there is some additional overlap between these ranges and the tick chart ranges. This would suggest that there was actually more agreement between the potential operational and pilot boundaries than the final outcomes would suggest, though the IQR ranges would have still suggested higher Y2 boundaries.

Table 22 *Operational and pilot judgemental grade boundaries – psychology 1*

	P1		P2		Overall	
Boundaries	A	E	A	E	A	E
Y1 operational	61	41	51	31	112	72
Y2 operational	55	32	47	28	102	60
Y2 RO	58	40	49	32	107	72
Y2 PCJ	57	39	53	32	111	71
Y2 t/c	52-58	30-35	44-51	26-31		
50% IQR RO	58-59	39-40	48-50	31-33		
50% IQR PCJ	56-59	38-40	52-55	30-33		
2SD RO	3	2	3	3		
2SD PCJ	5	3	5	5		

Table 23 shows that the views of paper difficulty differences were quite mixed and do not appear clearly related to the outcomes of the 2 pilots.

Table 23 *Judges' initial views of paper difficulty differences – psychology 1*

Pilot	Paper	Y1 more difficult	Y2 more difficult	Papers similar	Total
RO	P1	3	0	2	5
	P2	1	4	1	6
PCJ	P1	4	0	5	9
	P2	1	4	5	10

Taken together, the results of these pilots suggest that the CJ exercises succeeded in producing credible grade boundaries, based on a plausible script quality scale and high level of agreement between test score and quality measure scale. This is despite the fact that most boundaries, and especially at qualification level, are quite discrepant compared to the operational Y2 boundaries. The fact that the exercises largely replicated each others' grade boundary outcomes, with highly correlated script measures despite some of the judges not being the same, lends further support to the credibility of its outcomes, even where there were large discrepancies compared to the operational Y2 boundaries.

Pinpointing PCJ

According to initial mini RO estimates, Y2 grade boundaries were as follows:

Table 24 *Operational and mini RO judgemental grade boundaries – psychology 1 – pinpointing*

	P1		P2	
Boundaries	A	E	A	E
Y1 operational	61	41	51	31
Y2 operational	55	32	47	28
Y2 mini RO	58	40	49	31

Therefore, the following mark points were included in the pinpointing pilot:

- for paper 1 grade A: 56, 57, 58, 59, 60
- for paper 1 grade E: 38, 39, 40, 41, 42
- for paper 2 grade A: 47, 48, 49, 50, 51
- for paper 2 grade E: 29, 30, 31, 32, 33

Because the logistic regression analysis which was used to estimate the Y2 grade boundaries in this pilot does not rely on Rasch measures specifically but rather on the probability of whether a new session script beat the previous session script, we did not exclude those scripts that won or lost all their comparisons from the regression analysis. The grade boundaries for Y2 were calculated from the equations displayed in the table below.

Table 25 *Regression equations for calculation of Y2 grade boundaries – psychology 1 – pinpointing*

Boundary	Equation
P1_A	-1.38 + 0.02*x
P1_E	-6.04 + 0.15*x
P2_A	-4.30 + -0.08*x
P2_E	-3.50 + 0.11*x

Table 26 *Operational and pinpoint PCJ judgemental grade boundaries – psychology 1 – pinpointing*

	P1		P2		Overall15	
Boundaries	A	E	A	E	A	E
Y1 operational	61	41	51	31	112	72
Y2 operational	55	32	47	28	102	60
Y2 RO pilot	58	40	49	32	107	72
Y2 PCJ pilot	57	39	53	32	111	71
Y2 pinpoint PCJ	62	41	51	31	113	72

Taking into account the reliability levels achieved for each grade boundary, and the comparison of the grade boundary outcomes of pinpointing and the other 2 pilots, shown in the table above, we could probably place reasonable trust in the credibility of grade E boundaries for each paper. As for grade A boundaries, given low reliability levels, the fact that the resulting boundaries appear to align with those of the RO exercise might be to some extent reassuring, but not enough to conclude that the result for grade A from the pinpointing PCJ is uncontroversial.

Psychology 2

Table 27 presents the regression equations for calculation of marks, including grade boundary marks, in Y2 corresponding to equivalent performance in Y1. The mark-measure relationship is also presented in Figures 19 and 22, alongside the distribution of equivalent marks in Y2 corresponding to each of Y1 grade boundaries obtained through bootstrapping.

Table 27 *Regression equations for calculation of Y2 grade boundaries – psychology 2*

		Y	Equation	R2
RO	P1	Y1 mark	38.73+3.82*x	0.80
		Y2 mark	43.86+4.32*x	0.82
	P2	Y1 mark	40.10+3.79*x	0.87
		Y2 mark	40.36+4.17*x	0.92
PCJ	P1	Y1 mark	39.62+5.01*x	0.49
		Y2 mark	42.91+5.77*x	0.78
	P2	Y1 mark	40.61+6.38*x	0.78

[15] Weighting factor for paper 2 is 1.091.

Y2 mark	40.65+6.52*x	0.62

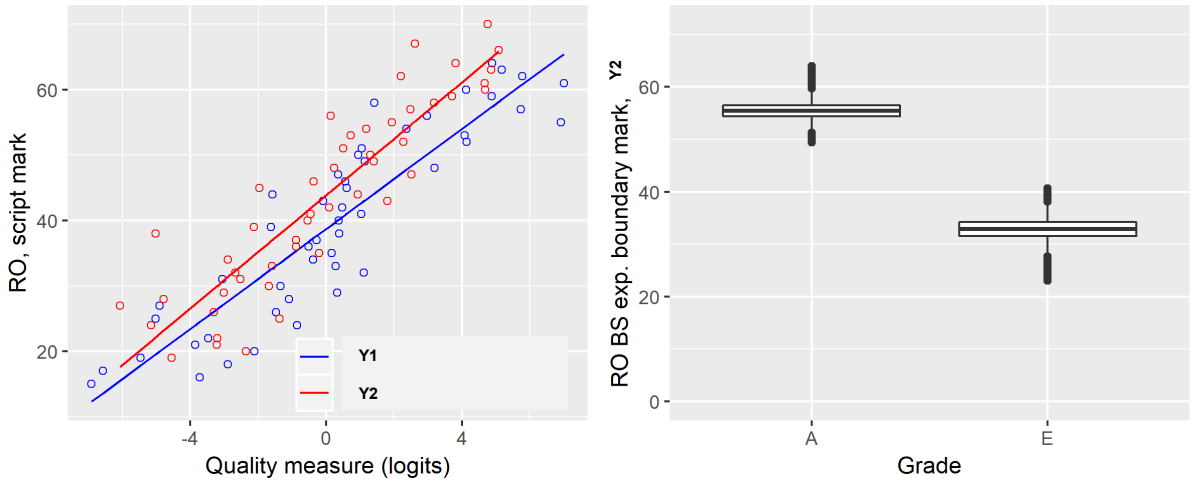


Figure 19 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – psychology 2 P1 – RO*

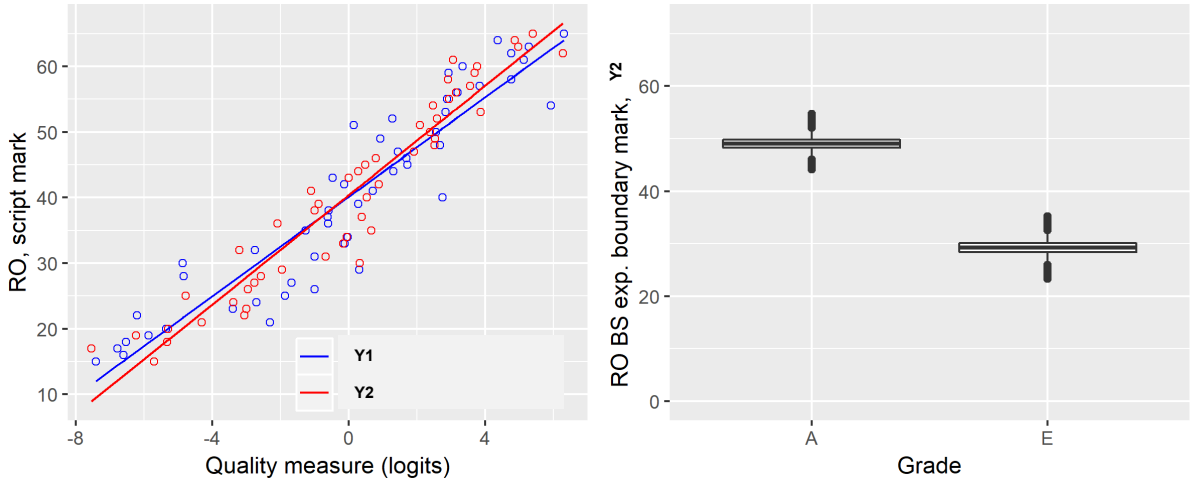


Figure 20 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – psychology 2 P1 – RO*

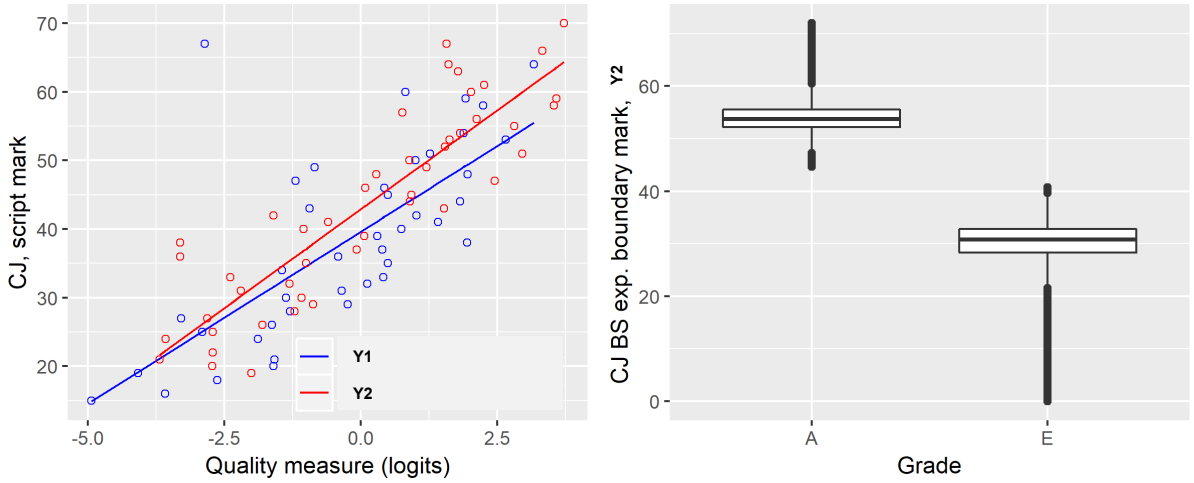


Figure 21 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – psychology 2 P1 – PCJ*

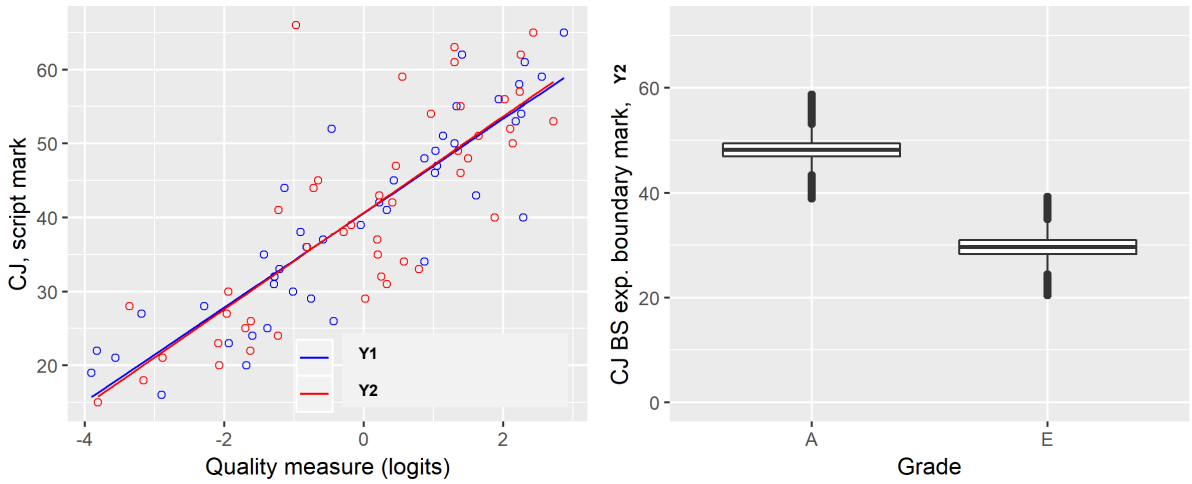


Figure 22 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – psychology 2 P2 – PCJ*

For RO, for paper 1, the interquartile ranges show that the middle 50% of marks corresponding to grade A boundary score fell in a narrow range of 2 marks. This was 3 marks for grade E. For Paper 2, the middle 50% range of marks was 2 for each grade.

For PCJ, for paper 1, the interquartile ranges show that the middle 50% of marks corresponding to grade A boundary score fell in a narrow range of 3 marks. This was 5 marks for grade E. For Paper 2, the middle 50% range of marks for grade A was 2, and 3 for grade E.

Table 28 shows that the RO pilot grade boundaries are within 2 marks of the Y2 operational ones for A boundary at paper level and qualification level. For E grade, the pilot boundaries are one mark away from operational boundaries. At qualification level, the pilot boundaries are identical to the operational ones.

The PCJ pilot grade boundaries for both papers and grades are very close to both operational and RO pilot A grade boundaries. Furthermore, the measures from RO and PCJ pilots were highly correlated at 0.90 for P1 and for 0.90 P2, suggesting that both exercises produced similar script quality scales in terms of script rank order.

The pilot boundaries were within the tick chart ranges in all but one case (PCJ paper 2 A boundary). However, the narrow IQRs largely overlapped in all cases between the pilots and with the tick chart ranges.

Table 28 *Operational and pilot judgemental grade boundaries – psychology 2*

	P1		P2		Overall	
Boundaries	A	E	A	E	A	E
Y1 operational	49	29	48	30	97	59
Y2 operational	53	32	51	30	104	62
Y2 RO	55	33	49	29	104	62
Y2 PCJ	55	33	48	30	104	63
Y2 t/c	51-55	30-34	49-53	28-32		
50% IQR RO	54-57	32-34	48-50	28-30		
50% IQR PCJ	52-56	28-33	47-49	28-31		
2SD RO	3	4	2	3		
2SD PCJ	5	8	4	4		

Table 29 shows that the views of paper difficulty differences were quite mixed and do not appear clearly related to the outcomes of the 2 pilots, although for paper 1, judges either thought that year 1 paper was more difficult or that they were similar. Only one judge thought that Y2 paper 1 was more difficult.

Table 29 *Judges' initial views of paper difficulty differences – psychology 1*

Pilot	Paper	Y1 more difficult	Y2 more difficult	Papers similar	Total
RO	P1	3	1	2	6
	P2	0	3	3	6
PCJ	P1	5	1	3	9
	P2	1	5	3	9

Overall, these pilots succeeded in producing credible grade boundaries, based on plausible script quality scales and high level of agreement between test score and quality measure scale. The fact that this exercise replicated the grade boundary outcomes of the RO exercise lends further support to the credibility of both outcomes.

English language 1

Table 30 presents the regression equations for calculation of marks, including grade boundary marks, in Y2 corresponding to equivalent performance in Y1. The mark-measure relationship is also presented in Figures 23 and 24 alongside the distribution of equivalent marks in Y2 corresponding to each of Y1 grade boundaries obtained through bootstrapping.

Table 30 *Regression equations for calculation of Y2 grade boundaries – English language 1*

	Y	Equation	R2
P1	Y1 mark	41.36+5.02*x	0.82
	Y2 mark	39.82+5.33*x	0.91
P2	Y1 mark	39.98+5.19*x	0.82
	Y2 mark	37.44+5.12*x	0.86

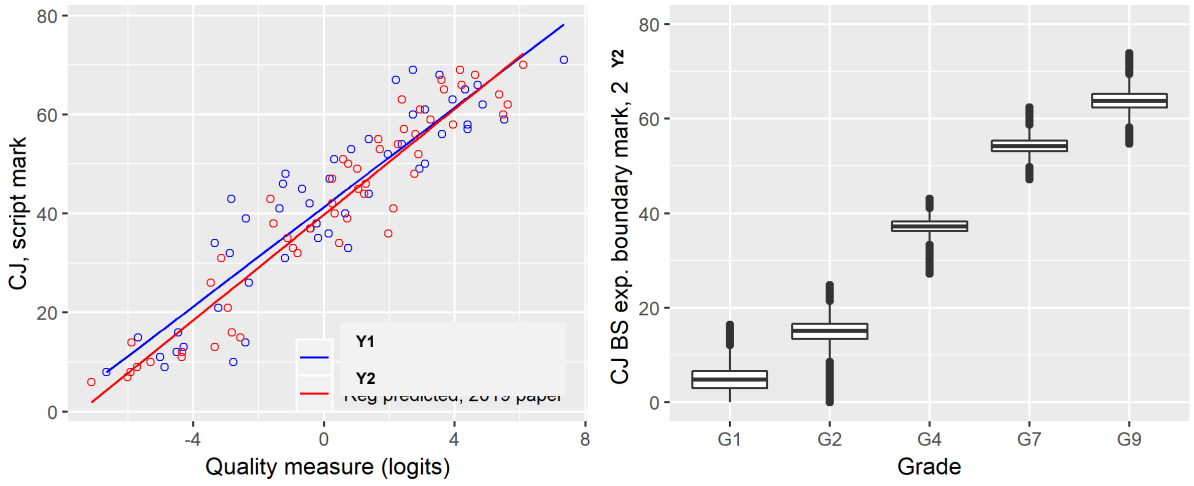


Figure 23 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – English language 1 P1*

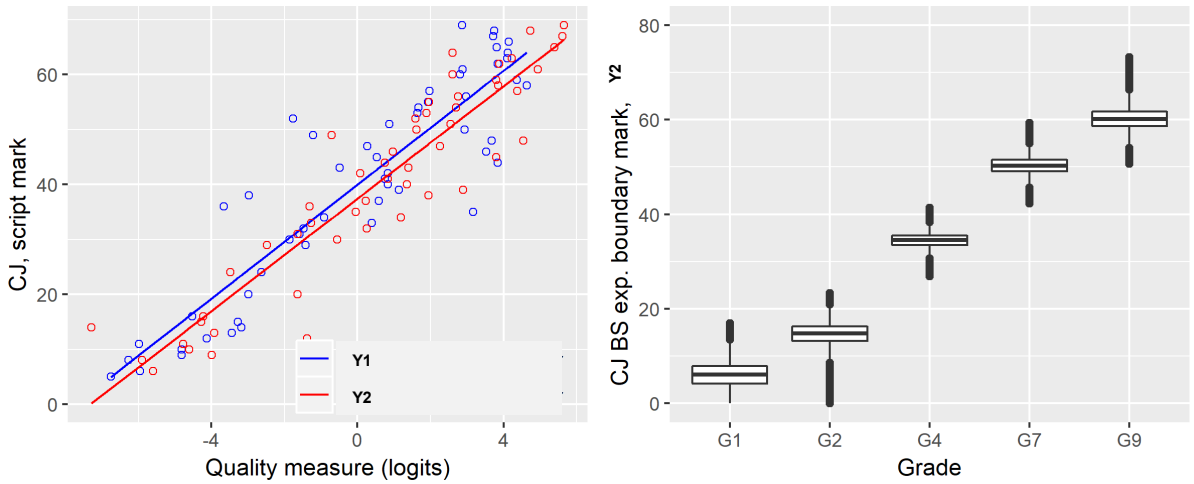


Figure 24 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – English language 1 P2*

The interquartile ranges show that the middle 50% of marks corresponding to grades 4 and 7 boundary marks fell in a narrow range of 2 marks. This was 4 marks for grade 1. It can be seen that there was somewhat more variability for grades 2 and 9 (3-mark range for the middle 50% of marks) than for grades 4 and 7, but there was less variability for grade 2 than grade 1 (3 vs. 4-mark ranges).

Tables 31 and 32 show that the pilot grade boundaries are reasonably close to operational Y2 boundaries, both at paper and qualification level, though lower for all 3 grades at qualification level and for most grades at paper level. Overall, this result would suggest that the Y2 papers were overall more difficult than the Y1 paper according to the PCJ result (across entire ability range for P2, and across lower ability range in P1).

The PCJ grade boundaries were within the tick chart ranges for P1 grades 4 and 7 and P2 grade 1, and just outside for the other grades. The bootstrapping exercise also suggests some potential variability in the PCJ boundaries, with the middle 50% ranges of 2-4 marks depending on grade boundary. It can be seen that there is some additional overlap between these ranges and the tick chart ranges, with only P2 grade 7 not having any overlap between the two. This would suggest that there was actually more agreement between the operational and pilot boundaries than the main estimates would suggest, though the IQR ranges would have still suggested slightly lower Y2 boundaries.

Table 31 *Paper level operational and pilot grade boundaries – English language 1*

	P1					P2				
Source	1	2	4	7	9	1	2	4	7	9
Y1	8	18	39	55	64	8	17	37	53	63
Y2	8	18	38	54	64	8	18	38	54	64
Y2 PCJ	4	15	37	54	64	6	15	35	50	60
Y2 t/c	6-10		36-40	52-56		6-10		36-40	52-56	
50% IQR	3-7	13-17	36-38	53-55	62-65	4-8	13-16	33-35	49-51	59-62
2SD	5	5	3	3	4	5	5	3	3	5

Table 32 *Qualification level operational and pilot grade boundaries – English language 1*

	Overall				
Source	1	2	4	7	9
Y1	16	36	76	108	128
Y2	16	36	76	108	127
Y2 PCJ	10	30	72	105	124

Table 33 shows that the views of paper difficulty differences were again quite mixed and not clearly related to the outcome of the CJ exercise in terms of paper difficulty differences.

Table 33 *Judges' initial views of paper difficulty differences – English language 1*

	Y1 more difficult	Y2 more difficult	Papers similar	Total
P1	5	5	5	15
P2	5	2	8	15

English language 2

Table 34 presents the regression equations for calculation of grade boundary marks in Y2 corresponding to equivalent performance in Y1. The mark-measure relationship is also presented in Figures 25 and 26 alongside the distribution of equivalent marks in Y2 corresponding to each of Y1 grade boundaries obtained through bootstrapping.

Table 34 *Regression equations for calculation of Y2 grade boundaries – English language 2*

	Y	Equation	R2
P1	Y1 mark	40.01+3.70*x	0.85
	Y2 mark	37.71+3.67*x	0.84
P2	Y1 mark	38.67+5.53*x	0.89
	Y2 mark	39.09+5.34*x	0.82

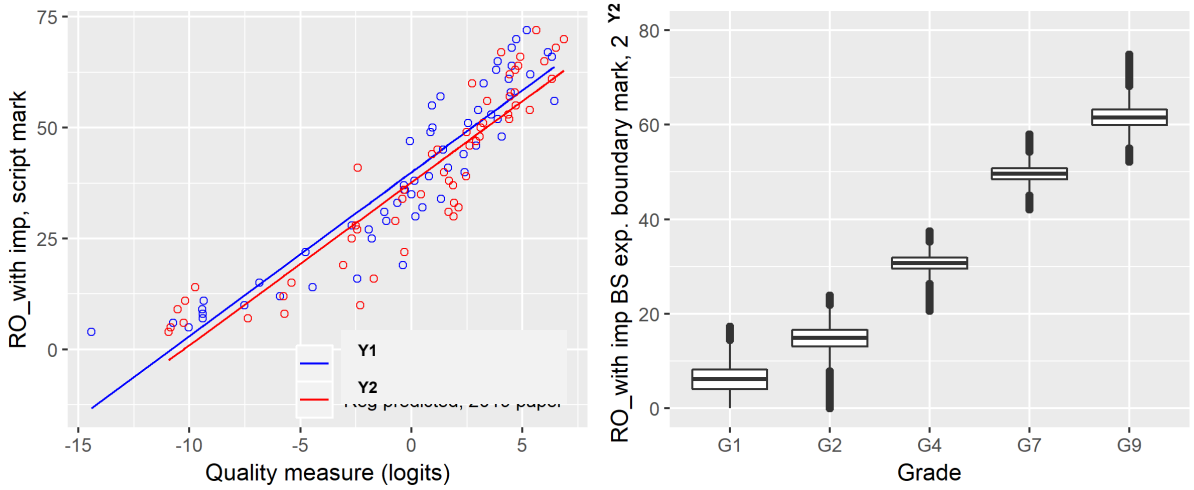


Figure 25 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – English language 2 P1*

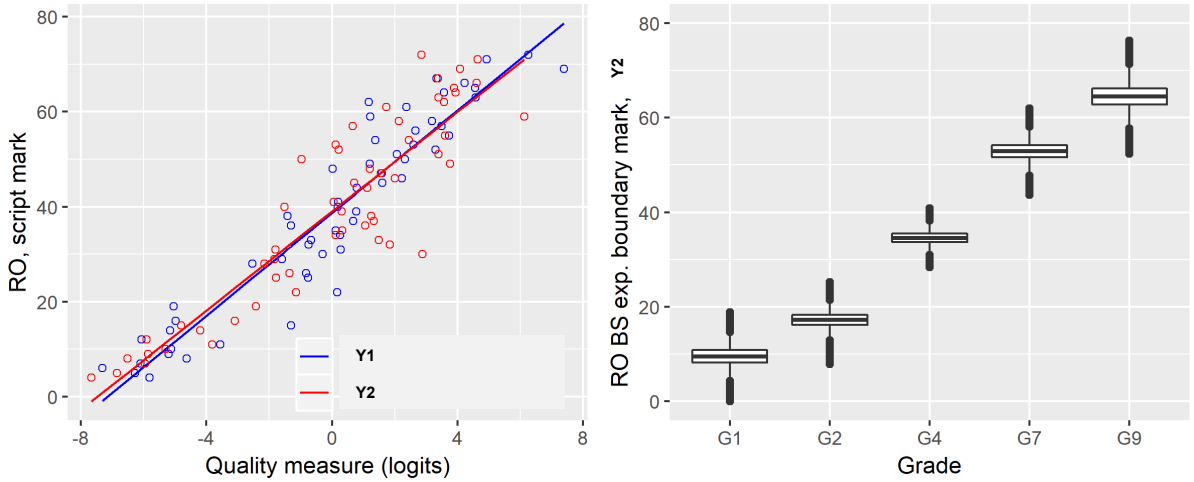


Figure 26 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – English language 2 P2*

For paper 1, the interquartile ranges show that the middle 50% of marks corresponding to grades 4 and 7 boundary marks fell in a narrow range of 2 marks. This was 4 marks for grade 1. For paper 2, the middle 50% range of marks was 2 for grade 4 and 3 for grades 1 and 7.

Again, there was somewhat more variability for grades 2 and 9 than for grades 4 and 7 (one-mark wider ranges for the middle 50% of marks), but there was less variability for grade 2 than grade 1 by one mark.

Tables 35 and 36 show that the pilot grade boundaries are reasonably close to operational Y2 boundaries, both at paper and qualification level. At qualification level, while Y2 operational boundaries were higher than Y1 operational boundaries, the pilot boundaries were all lower than both, and thus in the opposite direction from operational Y2 ones. This suggests that, according to the pilot results, the Y2 papers were more difficult than the Y1 papers overall. At paper level, the pilot boundaries also tended to be lower than Y2 boundaries, with the exception of P2 grade 1. The pilot grade boundaries were within the tick chart ranges in each case and the middle 50% bootstrap ranges also partially overlapped in each case. This overlap would suggest that there was actually more agreement between the operational and pilot boundaries than the main estimates would suggest, though the IQR ranges would have still suggested slightly lower Y2 boundaries in some cases.

Table 35 *Paper level operational and pilot grade boundaries – English language 2*

	P1					P2				
Source	1	2	4	7	9	1	2	4	7	9
Y1	8	17	33	52	64	8	16	34	53	65
Y2	8	16	34	53	65	8	18	36	54	66
Y2 RO	6	15	31	50	62	10	17	35	53	65
Y2 t/c	5-10		31-36	50-56		5-10		32-37	51-57	
50% IQR	4-8	13-17	30-32	48-51	60-63	8-11	16-18	34-35	52-54	63-66
2SD	6	5	3	3	5	4	3	3	4	5

Table 36 *Qualification level operational and pilot grade boundaries – English language 2*

	Overall				
Source	1	2	4	7	9
Y1	16	33	67	105	129
Y2	16	34	70	107	131
Y2 RO	16	32	65	103	126

Table 36 shows that the views of paper difficulty differences were again quite mixed and not clearly related to the outcome of the CJ exercise in terms of paper difficulty differences.

Table 37 *Judges' initial views of paper difficulty differences – English language 2*

	Y1 more difficult	Y2 more difficult	Papers similar	Total
P1	5	4	6	15
P2	7	0	7	14

English language 3

Table 38 presents the regression equations for calculation of grade boundary marks in Y2 corresponding to equivalent performance in Y1. The mark-measure relationship is also presented in Figures 27 and 28 alongside the distribution of equivalent marks in Y2 corresponding to each of Y1 grade boundaries obtained through bootstrapping.

Table 38 *Regression equations for calculation of Y2 grade boundaries – English language 3*

	Y	Equation	R2
P1	Y1 mark	31.65+4.61*x	0.86
	Y2 mark	32.84+3.97*x	0.91
P2	Y1 mark	38.69+5.41*x	0.89
	Y2 mark	39.06+5.22*x	0.82

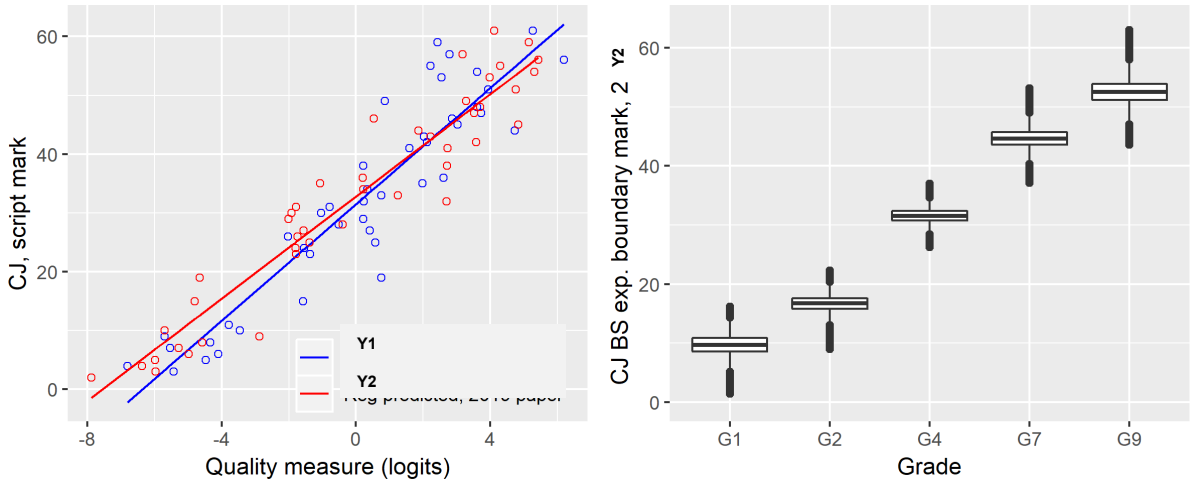


Figure 27 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – English language 3 P1*

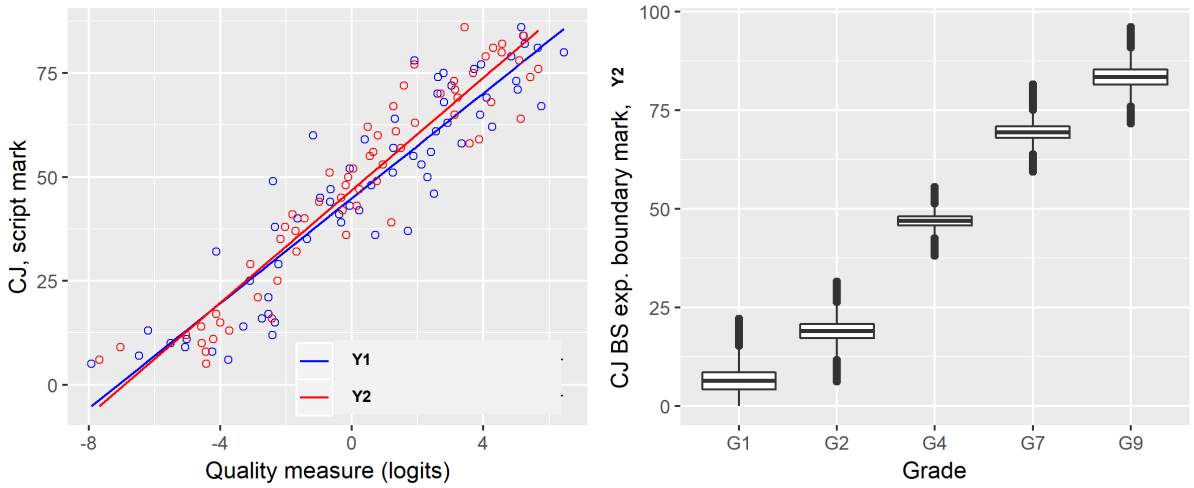


Figure 28 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – English language 3 P2*

For paper 1, the interquartile ranges show that the middle 50% of marks corresponding to all 3 key grade boundary marks fell in a narrow range of 2 marks. For paper 2, the middle 50% range of marks was 2 for grade 4 and 3 for grade 7 and 4 for grade 1.

In this case, the ranges for grades 1 and 2 were almost identical, while there was a bit more variability for grade 9 compared to grade 7 (one mark wider ranges for the middle 50% of marks).

Tables 39 and 40 that the pilot grade boundaries are reasonably close to operational Y2 boundaries, both at paper and qualification level. Except for grade 9, the pilot grade boundaries were lower than the operational Y2 boundaries. However, in this case, they were higher than the Y1 boundaries and thus in the same direction as the Y2 operational boundaries, suggesting that the Y2 paper was easier than the Y1 paper. The picture is more mixed at paper level, with some pilot boundaries lower, and some the same or higher than the Y2 operational boundaries.

The pilot grade boundaries were within the tick chart ranges in each case and the middle 50% bootstrap ranges also partially overlapped in each case. This overlap would suggest that there was actually more agreement between the operational and pilot boundaries than the main estimates would suggest, though the IQR ranges would have still suggested slightly lower Y2 boundaries in some cases.

Table 39 *Paper level operational and pilot grade boundaries – English language 3*

	P1					P2				
Source	1	2	4	7	9	1	2	4	7	9
Y1	5	13	30	45	54	7	19	45	66	79
Y2	9	17	34	47	55	11	23	49	69	80
Y2 PCJ	10	17	32	45	53	6	19	47	69	83
Y2 t/c	7-11		32-36	45-49		10-13		46-51	67-69	
50% IQR	9-11	16-18	31-32	44-46	51-54	4-9	17-21	46-48	68-71	81-85
2SD	3	3	2	3	4	6	5	3	4	6

Table 40 *Qualification level operational and pilot grade boundaries – English language 3*

	Overall				
Source	1	2	4	7	9
Y1	12	33	75	111	133
Y2	20	41	83	116	135
Y2 PCJ	16	36	78	114	136

Table 40 shows that the views of paper difficulty differences were again quite mixed and not clearly related to the outcome of the CJ exercise in terms of paper difficulty differences.

Table 41 *Judges' initial views of paper difficulty differences – English language 3*

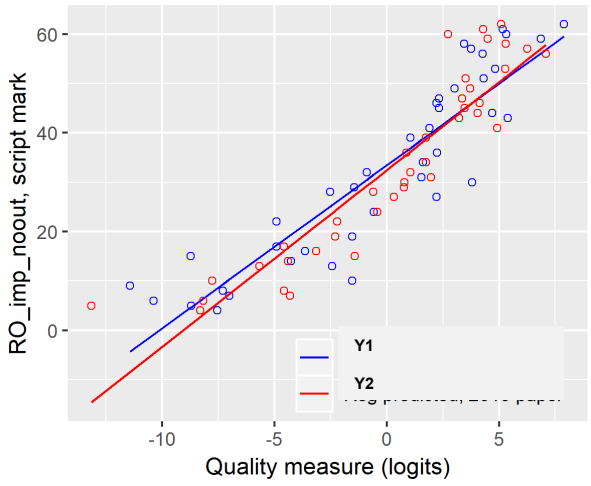
	Y1 more difficult	Y2 more difficult	Papers similar	Total
P1	5	5	9	19
P2	9	5	5	19

English language 4

Table 42 presents the regression equations for calculation of grade boundary marks in Y2 corresponding to equivalent performance in Y1 for the RO and PCJ exercises respectively. The mark-measure relationship is also presented in Figures 29 to 32 alongside the distribution of equivalent marks in Y2 corresponding to each of Y1 grade boundaries obtained through bootstrapping.

Table 42 *Regression equations for calculation of Y2 grade boundaries – English language 4*

		Y	Equation	R2
RO	P1	Y1 mark	33.47+3.30*x	0.82
		Y2 mark	32.45+3.58*x	0.84
	P2	Y1 mark	31.56+3.23*x	0.88
		Y2 mark	32.35+3.55*x	0.90
PCJ	P1	Y1 mark	33.76+4.77*x	0.90
		Y2 mark	32.04+5.14*x	0.91
	P2	Y1 mark	31.47+4.54*x	0.89
		Y2 mark	31.73+4.88*x	0.89



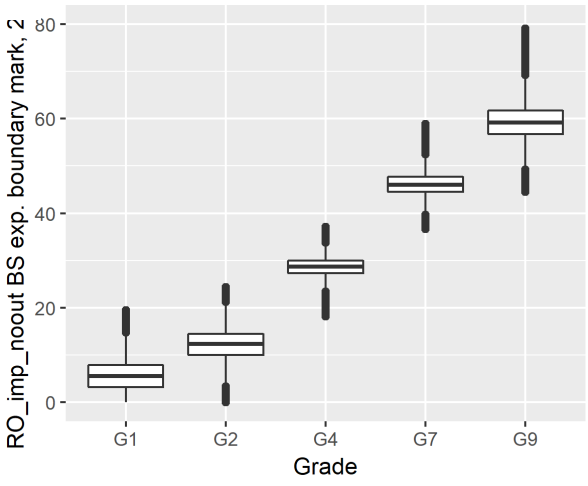


Figure 29 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – English language 4 P1 – RO*

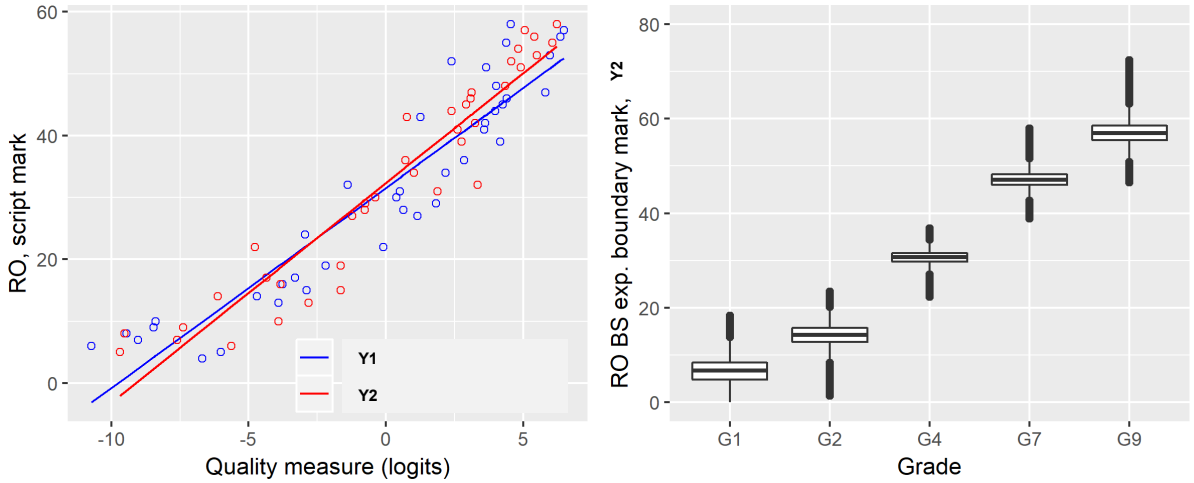


Figure 30 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – English language 4 P2 – RO*

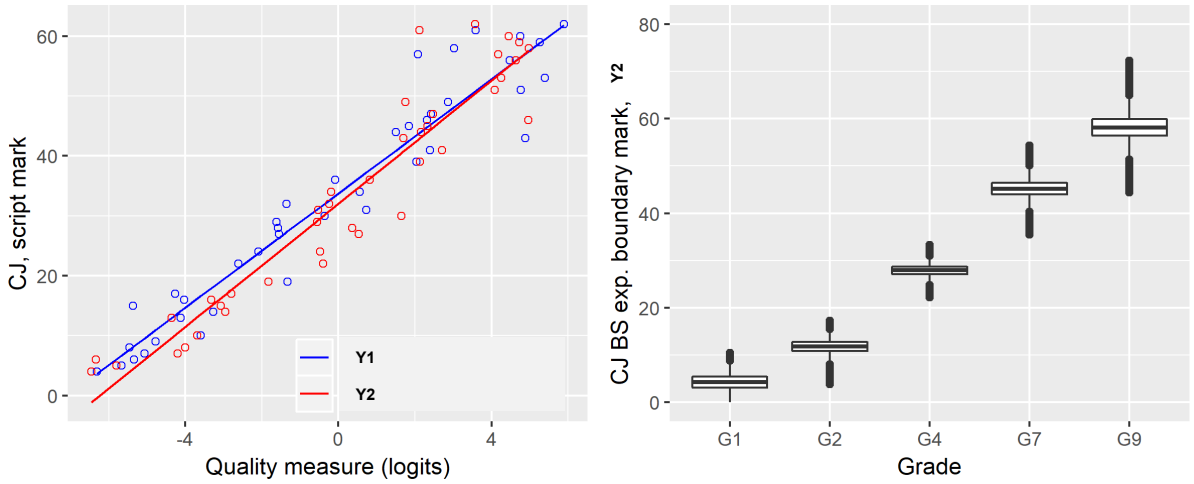


Figure 31 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – English language 4 P1 – PCJ*

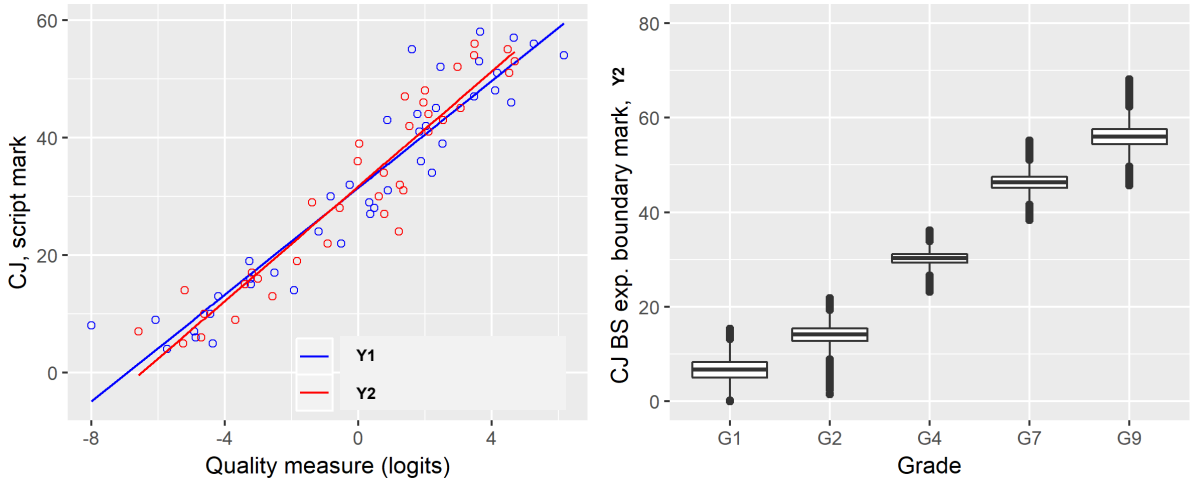


Figure 32 *Mark v. measure regression plot and bootstrap distributions of equivalent marks in Y2 corresponding to Y1 grade boundaries – English language 4 P2 – PCJ*

For the RO exercise, for paper 1, the interquartile ranges show that the middle 50% of marks corresponding to grades 4 and 7 boundary marks fell in a narrow range of 3 marks, while this was 5 marks for grade 1. For paper 2, the middle 50% range of marks was 2 for grade 4 and 7 and 4 for grade 1. Again, there was somewhat more variability for grades 2 and 9 than for grades 4 and 7 (1-2 marks wider ranges for the middle 50% of marks), but there was less variability for grade 2 than grade 1 by one mark.

For the PCJ exercise, for paper 1, the interquartile ranges show that the middle 50% of marks corresponding to all 3 key grade boundary marks fell in a narrow range of 2 marks. For paper 2, the middle 50% range of marks was 2 for grades 4 and 7 and 3 for grade 1. In this case, the ranges for grades 1 and 2 were almost identical, while there was a bit ore variability for grade 9 compared to grade 7 (one-mark wider ranges for the middle 50% of marks).

Firstly, Tables 43 and 44 show that the RO and PCJ boundaries were very similar to each other, though the PCJ qualification level boundaries were slightly lower than the RO boundaries. Across both of these pilots, the grade boundaries at both paper and qualification level were lower than the Y2 boundaries for grades 1 and 2, but higher or the same for the other boundaries. At qualification level, the pilot results would suggest that the Y2 papers were overall were slightly easier than the Y1 paper for higher ability students (grades 7 to 9) and slightly more difficult or similar for the lower ability students (up to grade 4).

Except for P1 grade 1 for both pilots, and P1 grade 1 RO pilot, the pilot grade boundaries were within the tick chart ranges in all other cases case and the middle 50% bootstrap ranges also partially overlapped in each case, both between the two pilots, and with the tick chart ranges. This overlap would suggest that there was actually more agreement between the operational and pilot boundaries than the main estimates would suggest. The fact that the 2 pilots agreed to a great extent lends additional credibility to the outcomes of these methods as, in each case, the judges were the same but they saw different sets of scripts.

Table 43 *Paper level operational and pilot grade boundaries – English language 4*[16]

	P1					P2				
Source	1	2	4	7	9	1	2	4	7	9
Y1	8	15	30	46	58	8	15	30	45	54
Y2	8	14	28	45	57	8	15	29	46	56
Y2 RO	5	12	29	46	59	6	14	31	47	57
Y2 PCJ	4	12	28	45	58	7	14	30	46	56
Y2 t/c	7-9		26-30	43-47		7-9		27-31	44-48	
50% IQR RO	3-8	10-14	27-30	45-48	57-62	5-8	13-16	30-32	46-48	55-59
50% IQR PCJ	3-5	11-13	27-29	44-46	56-60	5-8	13-15	29-31	45-48	54-58
2SD RO	6	7	4	5	8	5	5	3	3	5
2SD PCJ	3	3	2	4	5	5	4	3	4	5

Table 44 *Qualification level operational and pilot grade boundaries*

	Overall				
Source	1	2	4	7	9
Y1	20	38	75	114	139
Y2	20	37	72	114	141
Y2 RO	15	34	75	117	145
Y2 PCJ	14	33	73	115	142

Table 45 shows that the views of paper difficulty differences were again quite mixed and not clearly related to the outcome of the CJ exercise in terms of paper difficulty differences.

Table 45 *Judges' initial views of paper difficulty differences – English language 4*

	Y1 more difficult	Y2 more difficult	Papers similar	Total
P1	1	6	8	15
P2	2	2	11	15

Direction of grade boundary differences

Given that there is no way to disguise the exam session from which candidate performances came in a CJ exercise, it is conceivable that the judges may somehow be incentivised to judge the scripts from the more recent session as consistently better than those from the previous session in an attempt to ensure higher outcomes for candidates (lower grade boundaries). Even though it seems unlikely that this would be possible given the way the scripts were allocated to judges in CJ methods, we investigated the patterns with respect to the direction of grade boundary differences between pilot and operational boundaries across most of the pilots conducted (excluding pinpointing).

In order to do this, for each grade boundary (both at paper and qualification level) we calculated the difference between the pilot grade boundaries and operational boundaries for Y1 and Y2. We subtracted operational boundaries from the pilot boundaries, with positive difference showing when a pilot boundary was higher and negative difference when a pilot boundary was lower than an operational boundary.

[16] Note that weighting factor 1.5 is used for P2.

Firstly, as can be seen in Figure 33, there was a wide range of both positive and negative differences between pilot and Y1 (Diff Y1) and pilot and Y2 (Diff Y2) operational boundaries. Across all pilots and boundaries, a pilot boundary was equally likely to be higher or lower than either the corresponding operational Y1 or operational Y2 boundary, rather than, for instance, the pilot boundaries always being lower than the corresponding Y1 or Y2 operational boundaries.

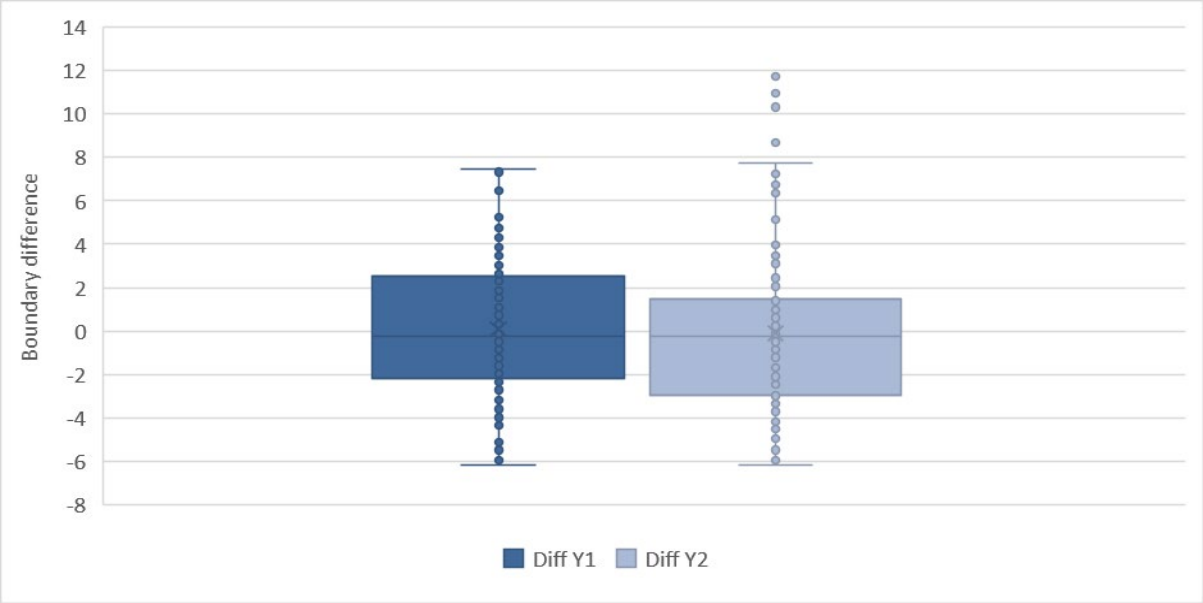


Figure 33 *Distribution of grade boundary differences between pilot and Y1/Y2 operational boundaries*

When broken down by method in Figure 34, the pattern is similar to the one above. The ranges of differences are wide and equally likely to be either positive or negative. This suggests that there is nothing inherently biasing in either PCJ or RO that might lead to predominantly lower or higher grade boundaries compared to either Y1 or Y2 operational boundaries.

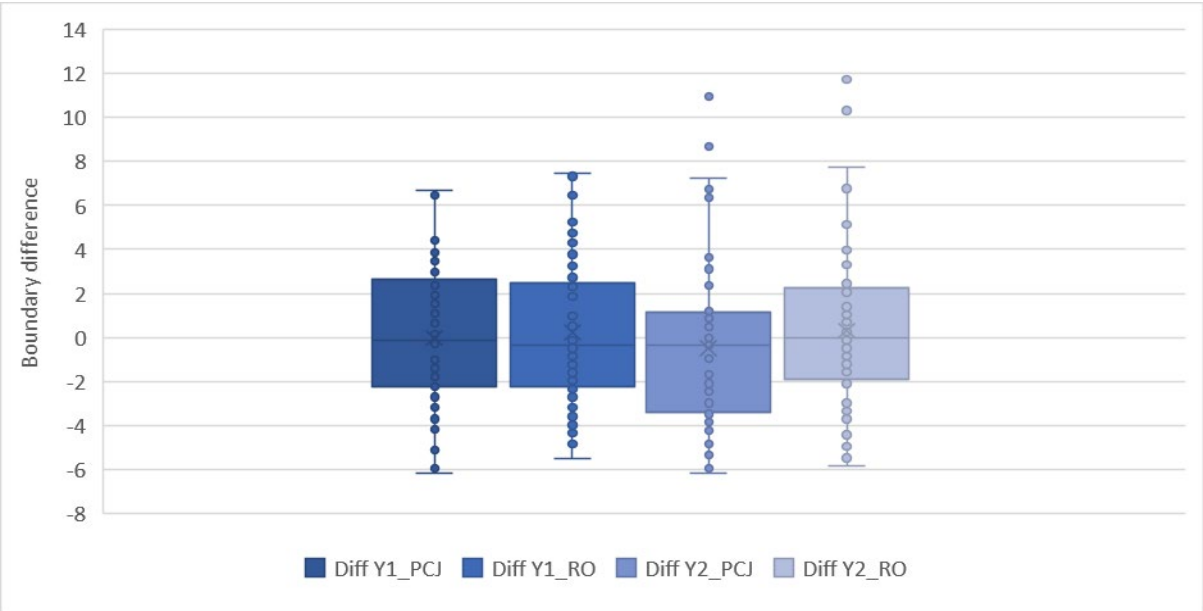


Figure 34 *Distribution of grade boundary differences between pilot and Y1/Y2 operational boundaries by method*

A very similar pattern in terms of positive vs. negative differences is apparent when these are broken down by whether the boundaries were paper-level or qualification level as in Figure 35. The range of differences is slightly wider at paper level, but the pilot boundaries were more or less equally likely to be higher or lower irrespective of whether they were calculated at paper or qualification level.

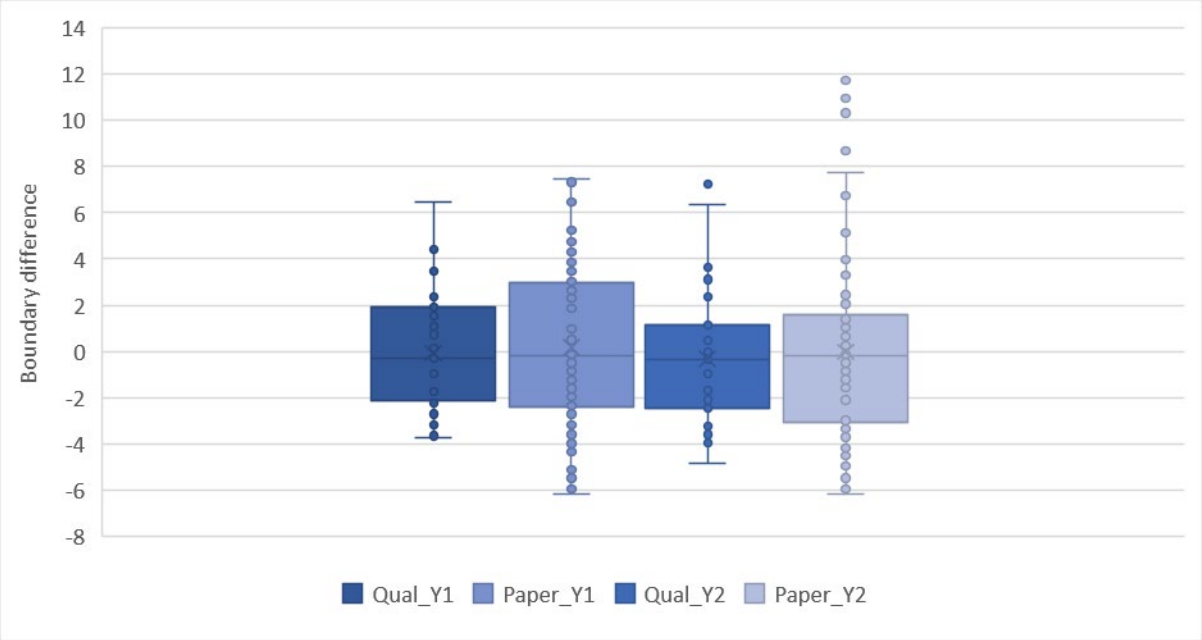


Figure 35 *Distribution of grade boundary differences between pilot and Y1/Y2 operational boundaries at paper and qualification level*

We also looked at the distribution of differences by grade boundary. The differences between pilot and Y1 boundaries are presented in figure 36, followed by differences between pilot and Y2 boundaries in figure 37. Note that boundaries 1 to 9 are all from English language pilots, while the A and E boundaries came from English literature and psychology pilots. As for comparison to Y1, we can see that for grades 1 and 2, and to some extent for grade 4, the pilot boundaries tended to be lower than the operational ones. For other grades, the differences were equally likely to be positive or negative. Comparing pilot boundaries to Y2 operational boundaries, we can see that the differences for grades 1 to 9 were mostly negative, i.e. the pilot boundaries tended to be lower. For the other 2 grades, the differences were again equally likely to be in either direction.

The pattern in English language (grades 1-9) below, indicates in most cases somewhat higher outcomes for candidates in Y2 (i.e. lower grade boundaries). This pattern may be subject-related, and may be the result of the stage of the reform English language is currently in, with some of the apparent increases in outcomes suggested by the CJ methods related to performance recovery after the introduction of new specifications.

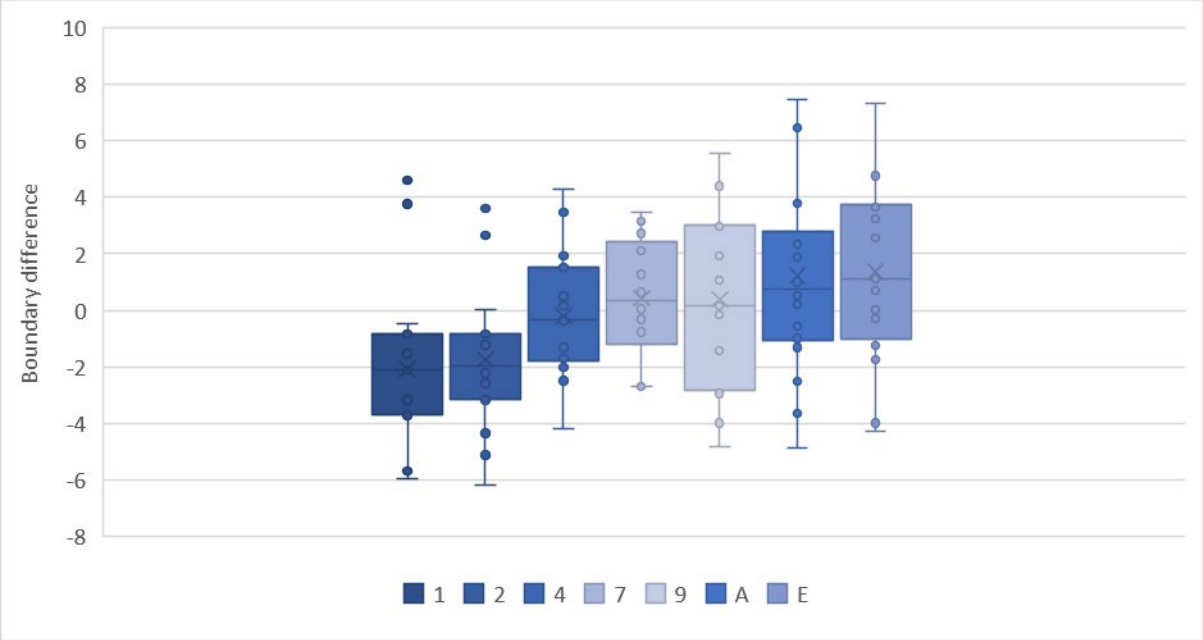


Figure 36 *Distribution of differences between pilot and Y1 operational boundaries by grade boundary*

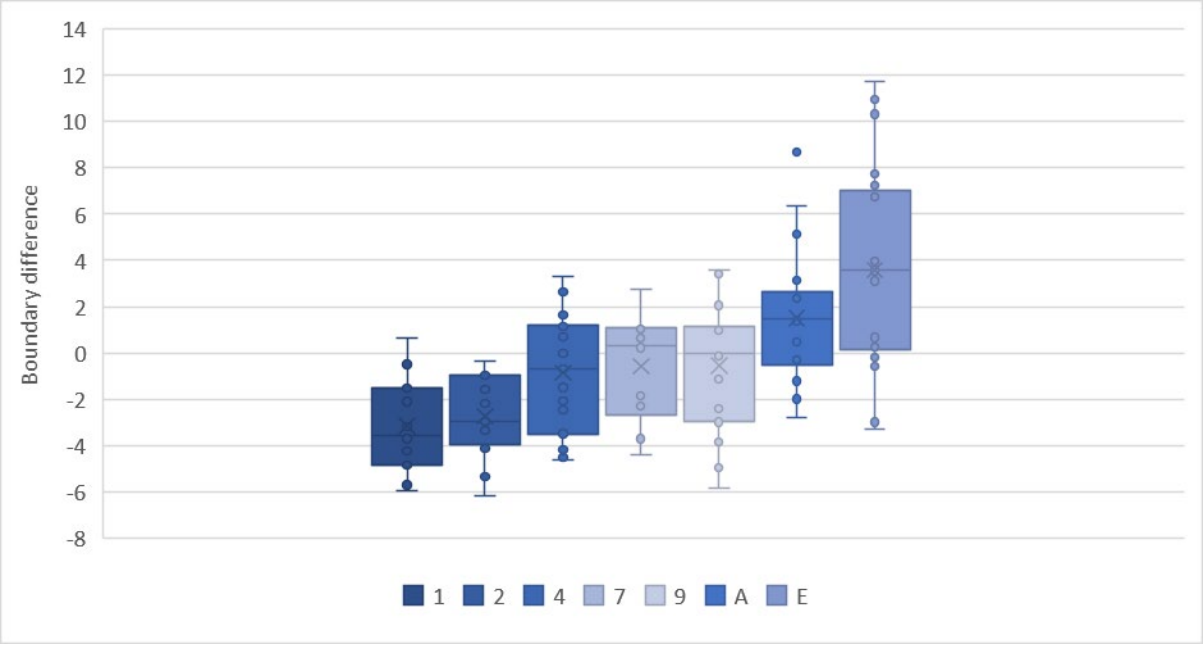


Figure 37 *Distribution of differences between pilot and Y2 operational boundaries by grade boundary*

Finally, looking at differences by individual specification (paper and qualification) in Figure 38, we can see that, while there is a pattern of lower pilot boundaries compared to Y2 operational ones for English language, there is no discernible pattern for English literature, media studies or psychology.

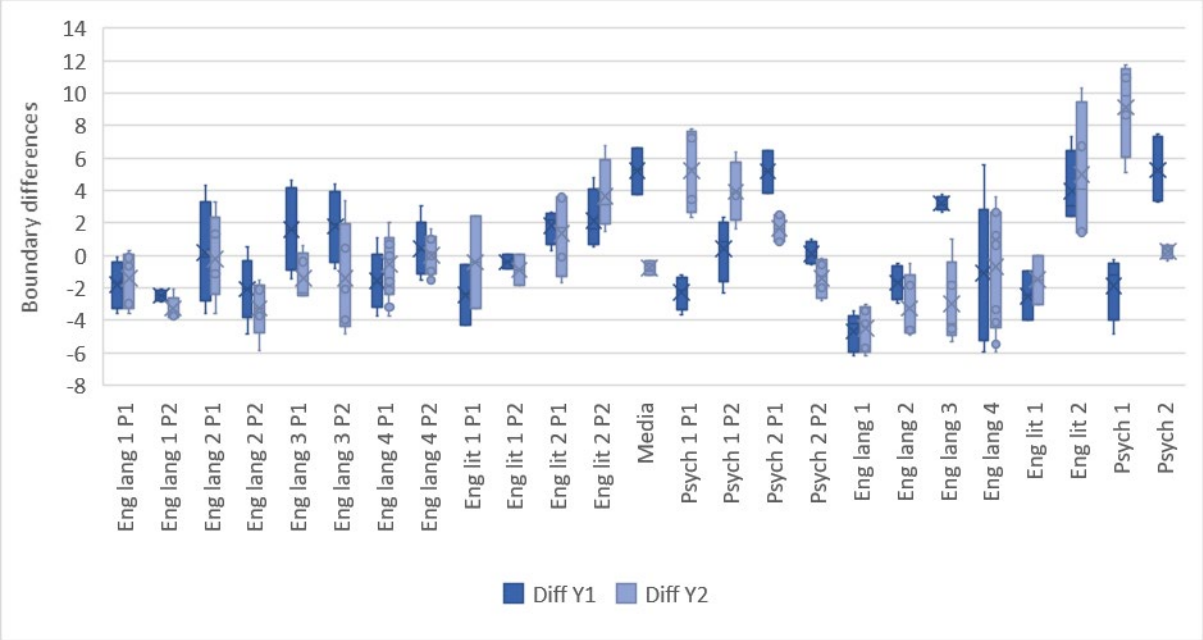


Figure 38 *Distribution of differences by specification (paper and qualification)*

These results suggest overall that there were unlikely to have been obvious method-related, boundary-related or specification-related effects with respect to the direction of pilot boundary differences, as these were generally equally likely to be higher or lower than the operational boundaries.

Effect of changing some design features and analytical decisions on the outcomes of CJ methods

In this section, we present the findings from analyses conducted on different subsets of the data collected for English language pilots. The size of these data sets allowed some leeway for data stripping and looking at outcomes based on different sections of the data. In particular, we investigated the effects of different judge group compositions (i.e. judge expertise in this case) and of reducing the number of comparisons per script. We evaluated the effects on scale properties and grade boundary outcomes. In addition, we investigated the inclusion or not of imputed measures in estimation of grade boundaries, and considered the effects this has on mark-measure correlations as well as specific grade boundary outcomes.

Judge expertise

The requirement for participation in our pilots was a minimum of 2 years examining experience in at least 1 paper from a particular specification. Typically, each group of judges assessing a specification consisted of some ordinary examiners with 2 or more years' experience, as well as some senior examiners, including team leaders, assistant principal examiners, principal examiners and chairs of examiners. Unlike our judge groups, the awarding panels currently only consist of senior examiners. It is conceivable that senior examiners may also be expected to perform better (or in some way differently) compared to ordinary examiners in CJ exercises based on their more extensive experience of marking in an operational context.

Given that senior examiners are typically more experienced, and possibly more reliable on average than ordinary markers, we explored whether the results of the

pilots would have differed substantially if judged by ordinary examiners vs. senior examiners. To do this, we split the groups of judges according to whether they were ordinary or senior examiners. The splits differ somewhat between different specifications, depending on the original composition of each group. In each case, where a subset of judges with the same expertise level was used in analysis, these were selected randomly within the relevant group. In the mixed groups, there are typically equal numbers of ordinary and senior examiners.

It should be noted that splitting the data like this leads to reduction in the number of comparisons per scripts, so some of the variability in the outcomes could be attributable to that rather than genuine effects of judge expertise, even where the groups are the same size. Nevertheless it may be useful to directly compare some of the groups to get a sense of the scale of variability in outcomes. If variability is found not to be substantial between ordinary and senior examiner groups, this would suggest that it may be fine to rely on ordinary examiners in CJ exercises. In operational live marking periods this could free up the time of senior examiners to engage in other responsibilities in relation to marking quality assurance.

As can be seen from the tables below, in general, there was little change in the grade boundary outcomes in different examiner groups, especially for grades 4, 7 and 9. There was somewhat more variability, as expected, for grades 1 and 2. In terms of direct comparisons between different groups of judges where there were equivalent number of comparisons per script, it can be seen that there is no consistent pattern in that more senior groups sometimes produced scales of higher reliability (e.g., English language 1, English language 2), and sometimes of lower reliability (e.g., English language 3, English language 4 RO and PCJ) compared to ordinary examiners or mixed groups. Furthermore, these differences are very small in all cases. This suggests that involving ordinary examiners in CJ exercises may be appropriate as their judgements seem comparable in quality and outcomes to those of their more senior colleagues.

Table 46 *Judge expertise effects – English language 1 P1*[17]

Judge group	N judges	Avg N comps	SSR	Sep	Corr Y1	Corr Y2	1	2	4	7	9
all	15	20	0.93	3.73	0.91	0.95	4	15	37	54	64
TLs	10	13	0.87	2.81	0.88	0.90	7	17	38	54	63
mix	10	13	0.88	2.85	0.86	0.94	1	13	37	55	66

Table 47 *Judge expertise effects – English language 1 P2*

Judge group	N judges	Avg N comps	SSR	Sep	Corr Y1	Corr Y2	1	2	4	7	9
all	15	20	0.93	3.73	0.91	0.93	6	15	35	50	60
TLs	10	13	0.86	2.68	0.89	0.90	6	14	33	48	58
mix	10	13	0.88	2.85	0.86	0.91	6	15	35	51	61

Table 48 *Judge expertise effects – English language 2 P2*

Judge group	N judges	Avg N comps	SSR	Sep	Corr Y1	Corr Y2	1	2	4	7	9
all	15	36	0.98	6.48	0.94	0.90	10	17	35	53	65
senior	9	22	0.97	5.77	0.96	0.92	8	16	35	54	67
senior	6	14	0.94	4.27	0.94	0.88	8	16	34	54	66
ordinary	6	14	0.93	3.66	0.88	0.90	12	20	36	54	65
mix	6	14	0.92	3.49	0.88	0.86	12	19	36	53	64

Table 49 *Judge expertise effects – English language 3 P2*

Judge group	N judges	Avg N comps	SSR	Sep	Corr Y1	Corr Y2	1	2	4	7	9
all	20	25	0.95	4.34	0.92	0.94	6	19	47	69	83
ordinary	15	19	0.93	3.76	0.92	0.94	7	20	48	70	84
mix	15	19	0.92	3.58	0.90	0.92	3	17	47	71	86
TLs	5	6	0.67	1.75	0.76	0.80	8	20	46	68	81

Table 50 *Judge expertise effects – English language 4 P2 - RO*

Judge group	N judges	Avg N comps	SSR	Sep	Corr Y1	Corr Y2	1	2	4	7	9
all	15	22	0.97	6.06	0.94	0.95	6	14	31	47	57
seniors	6	10	0.92	3.48	0.93	0.91	9	16	31	46	55
ordinary	6	10	0.94	4.22	0.90	0.90	9	15	29	43	52
mix	6	10	0.92	3.47	0.81	0.85	8	15	30	45	53

[17] In this pilot, for both papers, 10 of 15 judges were team leaders (TLs). Of the remaining 5 judges, one was ordinary examiner, and the rest assistant principal examiners. Hence the mixed group contains 5 randomly chosen TLs, 4 assistant principals and one ordinary examiner, and could be seen as on average possibly having higher expertise than the TL group.

Table 51 *Judge expertise effects – English language 4 P2 - PCJ*

Judge group	N judges	Avg N comps	SSR	Sep	Corr Y1	Corr Y2	1	2	4	7	9
all	15	22	0.94	4.2	0.94	0.95	7	14	30	46	56
ordinary	9	15	0.91	3.31	0.93	0.91	6	13	30	46	56
mix	9	15	0.90	3.20	0.92	0.94	5	13	30	46	56
seniors	6	10	0.78	2.13	0.87	0.91	11	17	31	44	53
ordinary	6	10	0.85	2.58	0.91	0.88	8	15	30	46	55
mix	6	10	0.81	2.32	0.90	0.93	9	16	30	44	53

The correlations and grade boundaries presented in these tables were calculated after excluding the scripts which had imputed measures. In smaller data sets this was sometimes over 20 scripts. We discuss the effects of retaining or removing the scripts with imputed measures in a separate section below.

Where we have stripped the average number of comparisons to 6 (Engl lang 3), and restricted the judgements to just those of the 5 senior examiners (TLs) who took part in that exercise, the SSR and separation are too low to be considered trustworthy and the correlations borderline acceptable even though the grade boundaries themselves are not dissimilar from those obtained from larger data sets. In this case, at least, the expertise of the judges did not in and of itself compensate for the small scale of the data set, and hence does not lead to the results in which we can have full confidence. We consider the impact of the number of comparisons per script on the outcomes in the next sub section.

Number of comparisons per script

As already noted in the introductory sections, number of comparisons per script has been shown in previous research to be a good indicator of likely scale reliability resulting from CJ exercises. In the English language pilots, the minimum number of comparisons per script was 20, which resulted in high SSRs in all cases, as was expected. However, data collection on this scale can be prohibitive in certain contexts, and it is useful to explore the scale reliability and variability of the results based on smaller numbers of comparisons. Previous research suggests that fewer than 10 comparisons per script is unlikely to lead to SSRs much higher than 0.7. Arguably, SSRs on that scale, while acceptable in some contexts, may not be appropriate for all uses.

We ran separate analyses to explore the impact of reducing the number of comparisons in some of our pilot data sets. The results are presented in Table 52 below. In some cases, the reduced number of comparisons was achieved by randomly removing all of the judgements of a subset of judges. In other cases, we retained some judgements from all of the judges that took part in an exercise, and removed some for each judge. The latter approach was intended to allow investigation of the effects on the results being based on judges' earlier judgements compared to judges later judgements. It was possible to strip the data in this way because all of the judgements made in No More Marking software have a time stamp.

Table 52 *Average number of comparisons effects*[18]

Paper	Avg N comps	N judgs	SSR	Sep	Corr Y1	Corr Y2	1	2	4	7	9
EL 1 P2	20	15	0.93	3.73	0.91	0.95	4	15	37	54	64
	13	10	0.88	2.85	0.86	0.94	1	13	37	55	66
	10	8	0.78	2.16	0.84	0.91	5	15	35	52	62
	10 (early)	15	0.83	2.39	0.87	0.87	6	15	33	48	57
	10 (late)	15	0.79	2.12	0.87	0.90	3	13	35	52	63
Simul. range							5	2	4	7	8
50% IQR							4	3	2	2	3
EL 2 P2	36	15	0.98	6.48	0.94	0.90	10	17	35	53	65
	14	6	0.92	3.49	0.88	0.86	12	19	36	53	64
	14	6	0.92	3.49	0.88	0.86	12	19	36	53	64
	10	6	0.87	2.81	0.76	0.84	12	19	35	52	63
Simul. range							3	2	1	1	1
50% IQR							3	2	2	3	3
EL 3 P2	25	20	0.95	4.34	0.92	0.94	6	19	47	69	83
	19	15	0.92	3.58	0.9	0.92	3	17	47	71	86
	10	8	0.80	2.26	0.82	0.89	7	20	47	69	83
	10 (early)	20	0.83	2.46	0.88	0.82	13	24	47	66	78
	10 (late)	20	0.84	2.50	0.85	0.89	7	20	47	69	83
Simul. range							10	7	1	4	8
50% IQR							4	4	2	3	4
EL 4 P2_PCJ	22	15	0.94	4.20	0.94	0.95	7	14	30	46	56
	15	9	0.90	3.20	0.92	0.94	5	13	30	46	56
	10 (early)	15	0.79	2.16	0.88	0.90	7	14	31	47	56
	10 (late)	15	0.83	2.41	0.93	0.90	6	13	28	44	53
	10	6	0.81	2.32	0.90	0.93	9	16	30	44	53
Simul. range							4	3	2	3	4
50% IQR							3	3	2	2	3

The most sensitive indicator of scale reliability change with reducing number of judgements per script is the separation index (Sep in the table). It can be seen that it tends to drop to just above 2 (though not in each case) when the number of comparisons per script is reduced to 10. SSRs also begin to drop below 0.8 with reducing number of comparisons per script. The results also show that the drop in reliability tends to be accompanied by a drop in mark-measure correlations. Reduction of the number of comparisons per script also tended to lead to an increase in variability in estimated grade boundaries, particularly at the extremes

[18] In the table, 'Simul. range' is the range of grade boundary differences from the simulations presented in the table for each subject. The '50% IQR' is the middle 50% inter quartile range for the relevant papers and pilots obtained through bootstrapping (presented also in the relevant tables in the Grade boundary results section).

(Simul. range in the table) compared to variability resulting from original, larger scale exercises (50% IQR), though not in all cases.[19]

It is conceivable that the benefit of a larger number of judgements per script may also be in the overall larger number of judgements that each judge has to make, which could lead to better (more consistent) judgments due to practice or memory effects. On the other hand, making a large number of judgements can also be tedious, which could impact negatively on judge reliability later in the judging window. It might be possible to get some insight into this by comparing the results based on early vs. later comparisons made by each judge (based on the time stamp when the judgement was made). An initial exploration of this in 3 papers did not show a consistent pattern, with reliability higher for earlier judgements in EL1 and EL3, but slightly lower in EL4.

We have not explored reducing the number of comparisons below 10 extensively, but see Table 49, where reduction to 6 comparisons per script led to a substantial drop in SSR and separation index. Furthermore, some of the PCJ pilots collected only 10-12 judgements per script, resulting in SSRs of 0.75-0.8 and some fairly low mark-measure correlations. Again, this would suggest that, unless SSRs of below 0.8 are deemed appropriate for a specific context, it would not be advisable to base the results of CJ exercises on fewer than 10 comparisons per script.

As in the previous section, the correlations and grade boundaries presented in the table above were calculated after excluding the scripts which had imputed measures. In smaller data sets this was sometimes over 20 scripts. We discuss the effects of retaining or removing the scripts with imputed measures in the next section.

Inclusion of imputed measures in estimation of grade boundaries

In the main analysis for each paper all scripts which won or lost all their comparisons, and hence had imputed measures and high associated standard errors, were removed from regression and mark-measure correlation analyses. However, while carrying out those analyses, it was apparent that whether imputed measures were kept or removed sometimes had a tangible impact on the final grade boundaries, and, to some extent, on mark-measure correlations.

For this reason, while carrying out the analyses to investigate the effects of some design features, we recorded the results both when imputed measures were retained and removed. These are presented in Table 53 below for the same groups of judges as in the previous section. The numbers in brackets in the first column represent the number of scripts with imputed measures that were removed from each data set.

As can be seen from the table, keeping or removing imputed measures has at least as much impact on grade boundary estimates and mark-measure correlations as reducing the number of comparisons, and sometimes more – as indicated by the grade boundary ranges based on removing imputed measures (Range_imp in the table) and the grade boundary ranges when reducing number of comparisons (Range_comps in the table). This would suggest that keeping or removing imputed

[19] The extent of variability would likely be larger in each case if bootstrap exercises were carried out for each case where the number of comparisons was reduced, as suggested by lower mark-measure correlations.

measures from these analyses should not be an arbitrary decision, and is something that needs justifying as well as reporting alongside the results.

It can also be seen that the number of imputed measures removed gets higher as the number of comparisons per script decreases. This suggests that if CJ exercises were run on a smaller scale, with fewer comparisons per script, it might be more likely for a larger number of scripts to win or lose all their comparisons, resulting in imputed measures. This would be another reason to avoid collecting too few comparisons per script where possible.

Table 53 *Effects of including or excluding imputed measures*

Paper	Imp	N judgs	N comps	Corr Y1	Corr Y2	1	2	4	7	9
EL1 P2	yes	15	20	0.91	0.95	4	15	37	54	64
(6)	no	15	20	0.92	0.94	8	16	35	50	59
	yes	10	13	0.88	0.93	8	16	35	51	60
(11)	no	10	13	0.86	0.91	6	15	35	51	61
	yes	8	10	0.84	0.91	4	14	34	51	61
(21)	no	8	10	0.84	0.91	5	15	35	52	62
	yes	15	10 (early)	0.87	0.91	8	16	34	49	58
(15)	no	15	10 (early)	0.87	0.87	6	15	33	48	57
	yes	15	10 (late)	0.90	0.92	5	14	34	50	60
(17)	no	15	10 (late)	0.87	0.9	3	13	35	52	63
Range_imp						5	3	4	6	7
Range_comps						5	2	4	7	8
EL2 P2	yes	15	36	0.94	0.89	13	20	35	51	62
(2)	no	15	36	0.94	0.90	10	17	35	53	65
	yes	6	14	0.87	0.86	10	18	34	52	63
(10)	no	6	14	0.88	0.86	12	19	36	53	64
	yes	6	10	0.75	0.8	10	18	35	53	65
(14)	no	6	10	0.76	0.84	12	19	35	52	63
Range_imp						3	2	1	2	3
Range_comps						3	2	1	1	1
EL3 P2	yes	20	25	0.93	0.93	8	21	47	69	82
(5)	no	20	25	0.92	0.94	6	19	47	69	83
	yes	15	19	0.91	0.94	2	16	46	71	86
(9)	no	15	19	0.90	0.92	3	17	47	71	86
	yes	8	10	0.86	0.90	7	20	47	69	83
(21)	no	8	10	0.82	0.89	4	18	48	72	87
	yes	20	10 (early)	0.89	0.87	5	18	46	69	83
(17)	no	20	10 (early)	0.88	0.82	13	24	47	66	78
	yes	20	10 (late)	0.87	0.89	8	20	46	67	80
(20)	no	20	10 (late)	0.85	0.89	7	20	47	69	83
Range_imp						11	8	2	6	9
Range_comps						10	7	1	4	8
EL4 P2_PCJ	yes	15	22	0.94	0.94	10	17	31	44	53
(4)	no	15	22	0.94	0.95	7	14	30	46	56
	yes	9	15	0.92	0.94	7	14	29	45	54

(8)	no	9	15	0.92	0.93	5	13	30	46	56
	yes	6	10	0.91	0.94	9	16	30	44	53
(15)	no	6	10	0.90	0.93	9	16	30	45	53
	yes	15	10 (early)	0.91	0.91	8	16	31	46	55
(17)	no	15	10 (early)	0.88	0.90	7	14	31	47	56
	yes	15	10 (late)	0.93	0.88	9	15	28	42	50
(8)	no	15	10 (late)	0.93	0.90	6	13	28	44	53
Range_imp						5	4	2	5	7
Range_comps						4	3	2	3	4

Judge survey results

Time taken per judgement

The No More Marking software calculates the median time each judge took to make their judgements. Below, we report the distribution of median time taken across PCJ pilots[20]. Figure 39 shows that the judges were taking 4 to 6 minutes on average to complete each comparison depending on specification and pilot. There was much more variation in this respect amongst teachers compared to the expert examiners involved in other pilots.

For RO, based on judges' own estimates reported in our surveys, they took on average half an hour to complete a pack of 6 scripts. This is the equivalent of about two and a half minutes per extrapolated comparison.

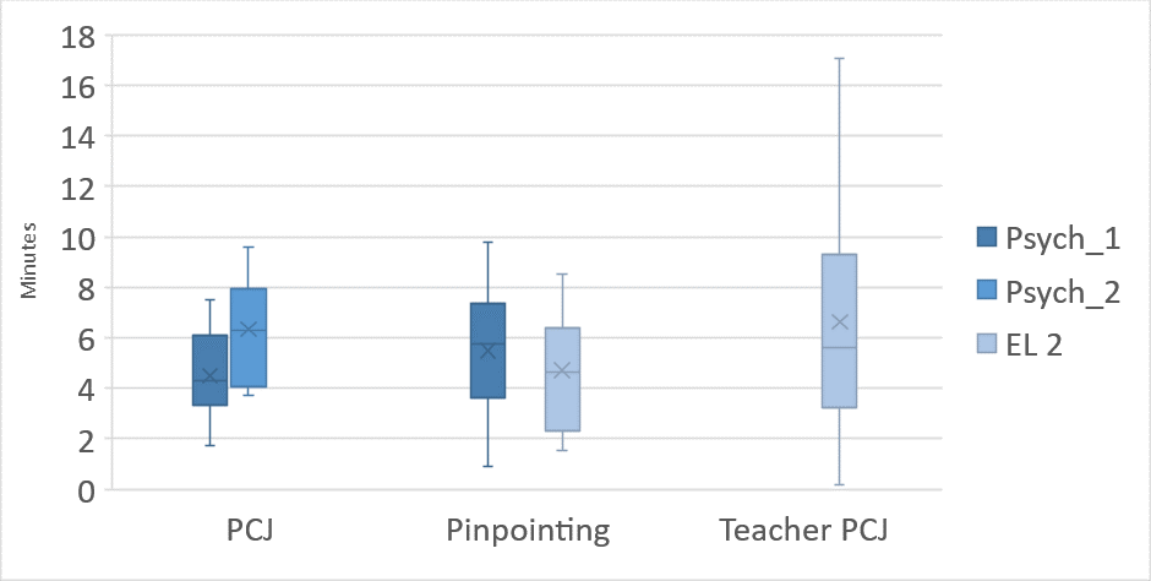


Figure 39 *Distribution of median time taken per judge by specification and method*

Task difficulty

Judges perceptions of task difficulty across different methods is shown in Figure 40. It can be seen that the majority of judges perceived the paired comparisons method as fairly easy, compared to RO, which was seen by the majority as fairly difficult.

[20] There were issues with time recording for English language pilots, and we therefore do not have reliable data on time taken for those.

However, given that the outcomes tended to be replicated where both RO and PCJ were conducted on the same scripts with mostly the same judges, this suggests that the judges perception of RO as more difficult did not result in substantially worse judgements.

The judges who did the pinpointing exercise were evenly split between seeing it as fairly difficult or fairly easy. Even though the pinpointing exercise was also done using online paired comparisons on No More Marking software, this suggests that it was seen as more difficult than the other online paired comparisons exercises. This is likely due to all the scripts involved in pinpointing being very close in terms of overall mark and presumably very similar in terms of perceived quality, making the judgements more difficult.

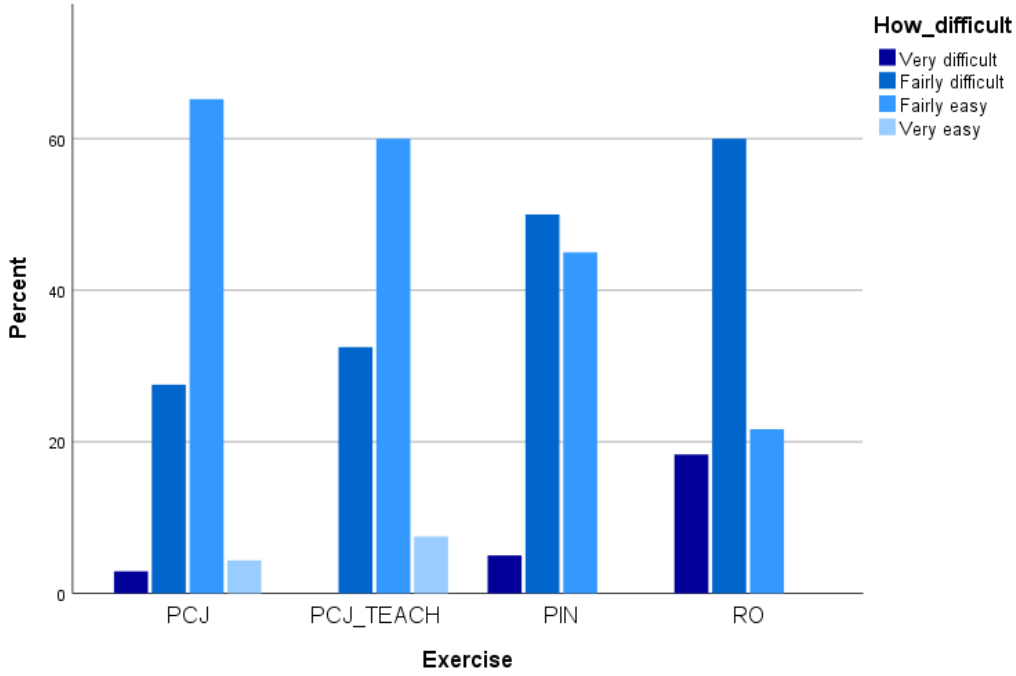


Figure 40 *Perceptions of task difficulty by method*

Did participation in live pilots interfere with marking responsibilities?

As already mentioned, some of the pilots were carried out during live marking and before awarding. For these pilots, because it was possible that some judges could drop out due to marking engagements which needed to take priority, or for any other reasons, we had arrangements in place to ensure they were still completed on time (i.e. before awarding). We thus had reserve judges on stand-by, we extended deadlines in some cases, and it was made clear to judges in their contracts and all communications that marking was to take priority over piloting deadlines.

The judges that took part in the pilots during live marking were asked in the survey following the pilots whether their participation interfered with their marking and other responsibilities during this time. As can be seen from the figure below, the judges overwhelmingly indicated that their participation in the pilots did not interfere with live marking.

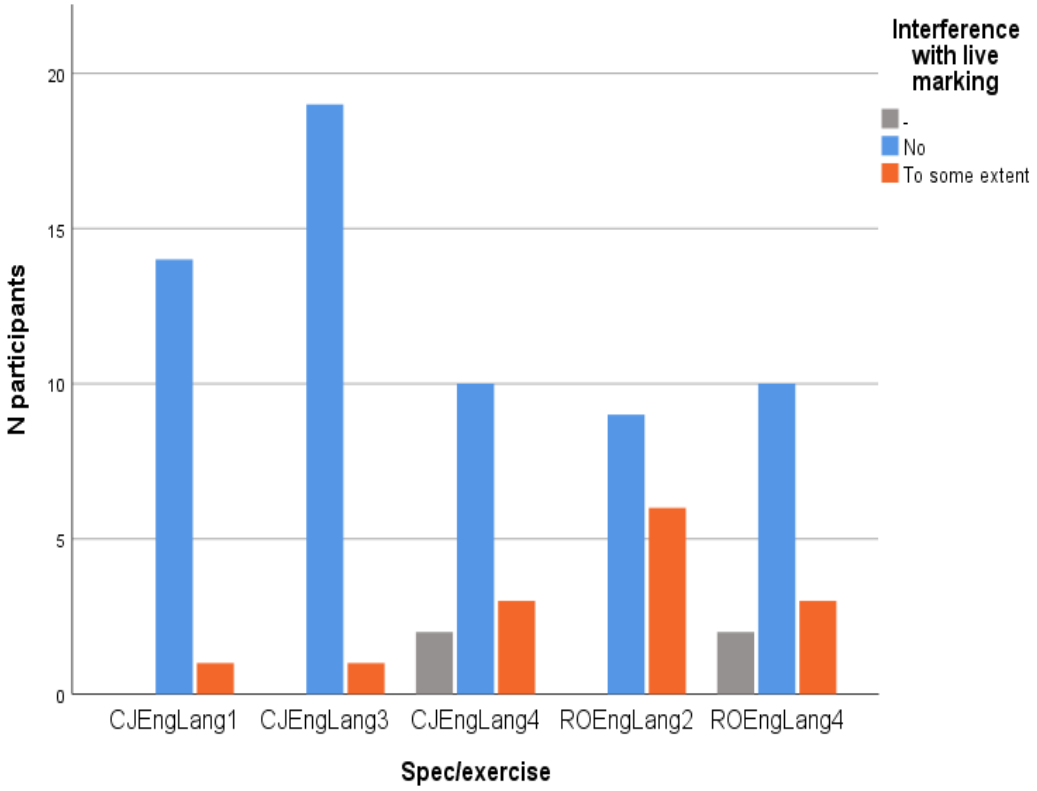


Figure 41 *Interference of CJ tasks with live marking*

Those that indicated that there was some interference were mostly senior examiners (team leaders and principal examiners), whose responsibilities with respect to marking quality control and stepping in to cover other markers' allocations towards the end of the marking window to some extent clashed with the time the pilots were taking place (which was also towards the end of the marking window). Some of their comments are shown below.

> *'I always work to make sure that I finish my marking quota early so that I can take on extra - to the benefit of both the examination board and myself. This year, I completed my allocation early - but this time it was in order to do these tasks; as a result, I have missed out on the extras this year.'*

> *'I completed all of my allocations early, but I did have a few issues with supporting my team to take on extra scripts that were left after the official deadline.'*

> *'I did devote several hours to the CJ task which I probably would have spent in adjudication, but I have completed my allocation.'*

> *'It meant that I had to stop marking whilst doing the study and will now need to work hard to meet the deadlines.'*

> *'I have finished my allocation but we are still finishing off so I have taken on extra scripts. It has meant that I have not been able to do as much marking over the past few days as I would normally have been able to complete.'*

How did judges account for differences in paper difficulty when judging script quality?

As part of the post-task survey for English language pilots, the judges were asked to state how confident they felt that they were able to differentiate between papers from different sessions in terms of difficulty and that they were able to take the differences in paper difficulty into account. As can be seen from Figures 42 and 43, the majority of judges for all specifications except for English language 2 stated that they were either very confident or fairly confident that they could do both things. However, as we have seen in the Grade boundary results section, their level of confidence did not seem to translate into their ability to determine paper difficulty differences that would consistently match the outcomes of the pilots.

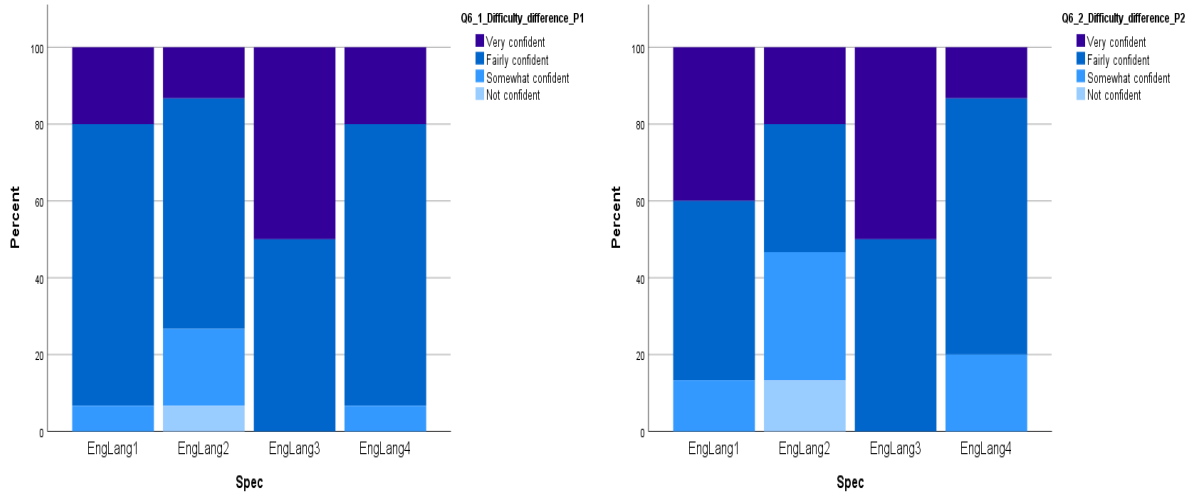


Figure 42 *Confidence in ability to differentiate between sessions in terms of difficulty*

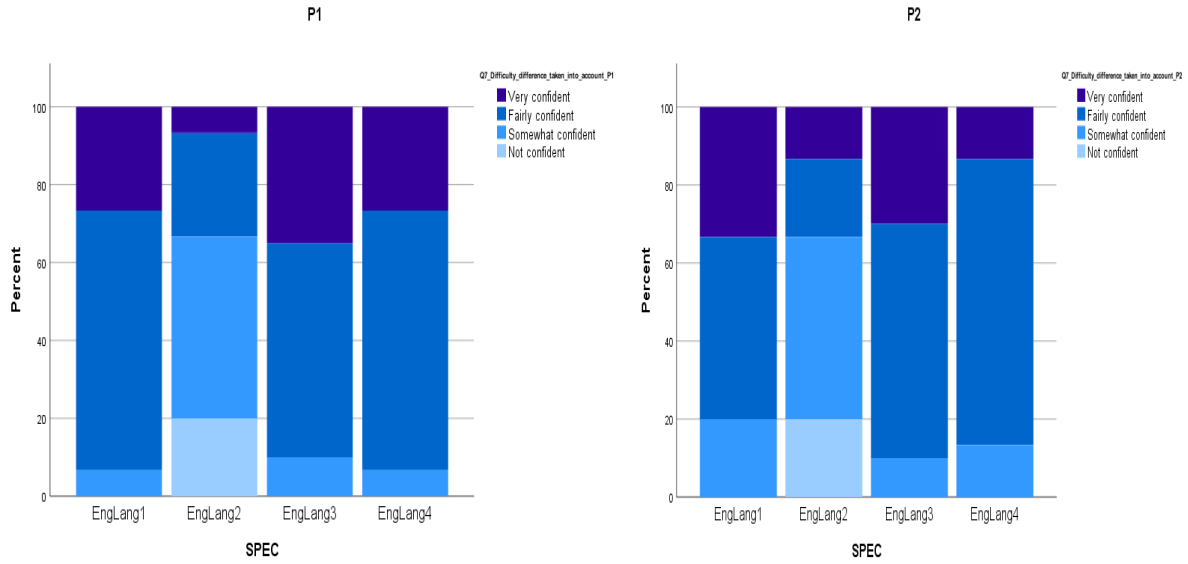


Figure 43 *Confidence in ability to take paper difficulty differences into account when judging quality*

The judges in other pilots were asked somewhat different questions about this. They were asked how easy or difficult they found it to take into account the differences between papers from different sessions when judging script quality. As can be seen in Figure 44, a larger proportion of judges across media studies and psychology felt that it was fairly difficult or very difficult to do this. The majority of English literature judges still felt that this was fairly easy or very easy. Again, the perception of ease as well as the perception of paper difficulty differences did not always relate to the direction of paper difficulty differences implied by the pilot outcomes.

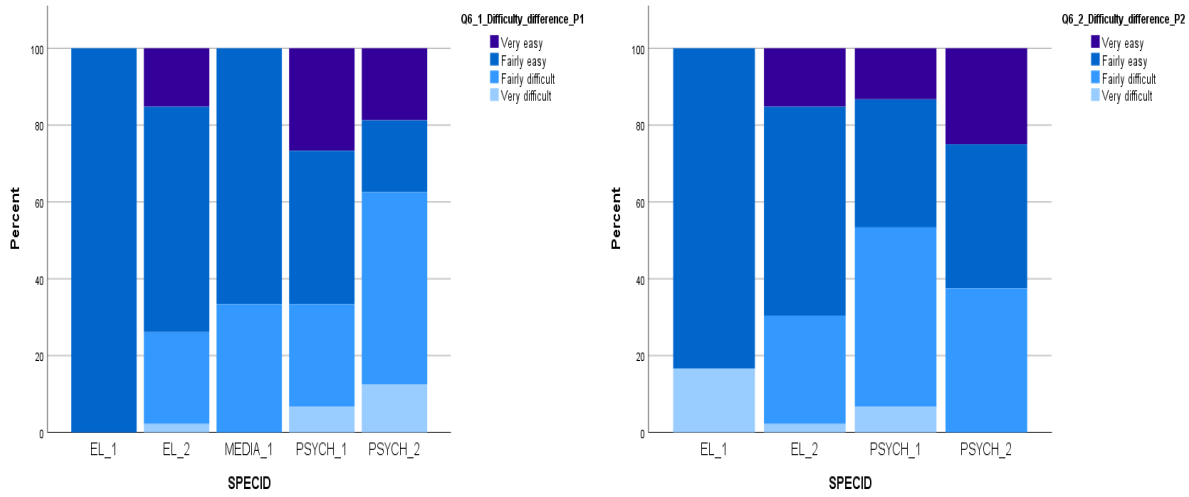


Figure 44 *Ease/difficulty of taking paper difficulty differences into account*

The judges were also asked to explain how they compensated for any perceived differences in paper difficulty between sessions. Below we present the analysis of the responses for English language specifications. To analyse the responses, a thematic analysis was performed using the QSR International's NVivo 11 software (2015).

Many examiners considered different sections of the papers or different questions separately when considering overall paper difficulty and consequent compensation for it. For this reason, a number of judges that judged the papers from different sessions to be similar in difficulty overall, still described a process of compensation in relation to individual sections or questions.

> *'For J351/1, I would take into account that the reading task for the Y2 paper was harder, and that the writing tasks were harder for the Y1 paper. If a candidate did not do so well in the writing section for the Y1 paper, but better in the reading section, I would put that below a candidate who did well in both sections of the Y2 paper.'* [Both papers of similar difficulty][21]

> *'I felt that where some questions might be more difficult, the other questions balanced this out, most specifically when it came to writing.'* [Y2 paper 1 more difficult, paper 2 similar between sessions]

[21] The comments in square brackets indicate the response the relevant judge gave when asked to compare papers from different sessions in terms of difficulty.

> *'I felt the reading section of the Y2 paper was easier than that of Y1.'* [Y1 paper 1 more difficult, paper 2 similar between sessions]

Having formed a view of relative paper difficulty between different sessions, the majority of examiners said they compensated for difficulty in first reading rather than the second. Whilst some examiners simply stated that they bore the paper difficulty in mind while judging, the majority went into a more detailed explanation of how they compensated for differences in difficulty between papers. For example, the majority stated that they adjusted their script judgements by either being 'more generous' or not 'overly harsh' when judging responses to a paper which they deemed more difficult.

> *'When presented with a very high standard response from each paper I might well judge the [Y1] one to be better.*' [Y1 more difficult for both papers]

> *'Q2 Y2 seemed more difficult in terms of what students could actually infer from Y1 and so I kept this in mind when comparing with Y2 and wasn't overly harsh.'* [Y2 paper 1 more difficult]

> *'For Q3, I felt there were fewer examples of obvious features and less familiarity with the type of text for the Y1 paper, so I adjusted my assessment in terms of being a little more generous'* [Y1 paper 2 more difficult]

A few examiners stated their compensation method was to change their expectations of candidates answers based on paper difficulty.

> *'I felt students had to spend longer reading and understanding the Y1 text before writing. Therefore I accepted that students often wrote shorter answers on the Y1 questions.'* [Y1 more difficult both papers]

> *'With the writing tasks, it seems easier to introduce speech-type features than magazine/online forum features, so I would have lower expectations of the latter tasks.'* [Y1 more difficult for both papers]

> *'I felt the paper 1 text was easier in Y2 so I expected slightly more inference and slightly more points to be discussed.'* [Y1 more difficult for both papers]

Some judges reported that they compensated in different ways for different ability levels, if the questions targeted at those levels differed in difficulty between sessions.

> *'[…] I gave more credit to lower ability candidates attempting the transactional writing on the Y1 paper as both tasks were quite challenging for them.'* [Y1 paper 1 more difficult]

> *'It was more of a balancing act with Q6 as I feel there were difference in terms of lower and more able students, but again it was lowering or raising expectations. For example, on the Y2 paper, I feel that it was easier for very low ability candidates to give an appropriate response to the question, so again I was a little more*

> *lenient at this low end on the Y1 paper.'* [Y1 paper 2 more difficult, paper 1 similar between sessions]

Other judges also mentioned that they could perceive difficulty differences for different ability levels, thought they did not explicitly explain how they went on to compensate for those.

> *'Paper 1 - Florence extract more accessible. Less able students were more likely to attempt these questions.'* [Y1 paper 1 more difficult]

> *'The texts for Y2 seem to be more accessible for all students for Paper 2. Firstly, in terms of content – just the difference between surfboards and boats seems to make the paper in Y2 more accessible for the lower students particularly.'* [Y1 paper 1 and 2 more difficult]

> *'You can visibly see the structure clearer than in the Y2 text, which uses lots of long and complex sentences, making it possibly difficult for less able students to comprehend.'* [Y2 paper 2 more difficult]

Alternatively, several examiners opted to compensate for paper difficulty by redirecting their judgement focus to aspects of the examinees' response that was less impacted by the differences in paper difficulty in an attempt to make the scripts more comparable across years without judging one script more or less harshly than the other.

> *'When deciding on which script was better with regards this question, I considered how the pupil had analysed the feelings of the characters overall – slightly ignoring the 'changing' feelings concept for the Y2 responses. This made the playing field a little more even.'* [Y1 paper 1 more difficult, paper 2 similar]

> *'The question on Y2 P1 about relationship was quite difficult when compared with the questions on Y1 so I looked for a more general understanding here rather than points made.'* [Y2 paper 1 more difficult]

> *'P1, Q5 for instance - was more accessible in Y2 than in Y1, so I took into account the candidates' ability to focus on the question asked rather than the quality of their writing.'* [Y2 paper 1 more difficult, paper 2 similar]

Overall, examiners seemed to focus on individual question or section difficulty differences rather than whole paper difficulty. They compensated for paper difficulty primarily by being more or less generous when considering script quality, and to some extent adjusting their expectations for different ability levels. Additionally, examiners seemed to adapt how they judged examinee scripts in terms of which content they focused on in order to make the scripts more comparable across years. Finally, they changed their expectations of examinee responses so they were different across years in accordance with their judgement of paper difficulty.

How did judges make holistic quality judgements?

The judges were also asked to explain their general approach to making judgements about relative quality of the scripts they were comparing. In addition, they were asked if there were any particular features of scripts or responses to particular questions that influenced their judgements. A thematic analysis was performed on the responses from the judges who took part in English language pilots to look at both general approach and response features considered. A further analysis of response features was conducted on the responses from the judges in the English literature pilots. This had a further aim of comparing the features considered by the examiners vs. the teachers. The analyses were conducted using the QSR International's NVivo 11 software (2015).

General approach in English language pilots

Several themes emerged with respect to the general approach that the judges took when making comparisons. Most prominently, the judges noted that they tended to look at questions separately, either as the main approach or in addition to also trying to make a holistic judgement. A number of judges said that they looked at reading and writing questions separately:

> *'I assessed the reading fist making notes on their achievements, then assessed the writing section'*

> *'I looked at the two parts of each paper - reading and writing - sometimes a student's reading response was stronger than the writing and vice versa to form the overall judgement.'*

> *'A holistic overall judgement- looking at all aspects of the paper. However, I would often read through a whole paper and then judge the reading and writing sections comparatively.'*

Some judges did not specifically mention whether they looked at reading or writing separately, but did note that they read and evaluated each question at a time before making the overall judgement.

> *'I read through the scripts, a question at a time, and tallied the more successful response to each question.'*

> *'I would then look at each paper separately and whilst trying not to precisely mark each question, develop a sense of overall quality of each response in the paper. When that still did not divide a pair, I took time to scan each paper again, question by question, Looking for any point at which one had the advantage. There are lots of routes to the same level of quality and some were very close.'*

> *'I approached the packs in a couple of different ways. Towards the end, I would tend to begin by looking at the writing task(s) to gain a sense of overall quality. I would then go through the longer questions one by one, followed by the shorter ones - moving the papers around until they sat in a rough rank order.'*

Several judges also noted that they considered responses in greater detail when they could not make an initial clear judgement based on the first reading or when they felt that there was a tie between two scripts.

> *'Sometimes it was close but the quality of the writing response led to the judgement. Sometimes I had to do some adding up to see which was better overall, when the script quality was close.'*

> *'For some pairs of scripts it was very easy to see which was better holistically; for others I had to read them through again in order to determine which one would have been better in terms of its content.'*

Judges commonly noted that they were more influenced by candidates' performance on the writing task as opposed to the reading task, due to its greater length and a higher 'weighting'. This was particularly apparent in cases where scripts were perceived to be of a similar quality.

> *'The writing section helped to clarify my judgement. Typically, the candidates would be roughly around the same for each of the reading questions, although P1, Q3 and P2, Q2 were sometimes the anomalies so I tried not to make judgements based on these questions alone.'*

> *'Questions 4 and 5 influenced my judgments more heavily, as they are weighted more heavily and provide candidates with the largest opportunity to demonstrate their skills.'*

> *'For both P1 and P2 the candidates' writing responses had an influence as in both papers this is the question with most 'weighting'.'*

Some judges noted more generally that their judgements were influenced mostly by higher tariff questions (not just writing).

> *'I put more value on questions 3 and 4 from the reading sections, 5 or 6 from the writing section, because they have the largest mark allocations.'*

> *'The higher mark questions obviously matter more than anything else on most scripts although the first and second question become increasingly important as we go lower down the quality scale.'*

> *'The quality of Q3 and Q4 and then the quality of the writing task for me were good indicators – although I did my upmost to mark holistically and these for me were indicators not hard and fast rules, sometimes the best of a script would come from Q2.'*

Another influential aspect was whether candidates attempted every question or missed some out. The papers with unanswered questions were often judged as a lower quality compared to full scripts, regardless of the content of their answers, as judges seemed to assume that those who at least attempted the question might gain some marks compared to those who did not.

> *'I also took into account whether all questions had been answered – it was obvious that where a candidate had only answered perhaps 3 questions that this would likely be on the lower end of the ranking.'*
>
> *'If questions were missed out or the answers were extremely short, then I judged that that reduced the quality of the script.'*
>
> *'If one answer was blank and the other had writing then as long as the writing was of at least one rewardable mark I would favour that paper.'*

Within this category, some responses suggest that missing responses may have been used to some extent as 'quick' differentiators between scripts irrespective of the detailed aspects of performance.

> *'I would say that upon first glance, the quantity of response definitely influenced my judgement In the sense that if someone did respond to the questions asked in comparison to another candidate who did not respond at all - it stands to reason that the first candidate would gain 'some' marks regardless of which band, compared to the second candidate who would receive none.'*
>
> *'If students had missed large questions that really hindered the overall mark. If students had missed random questions and the other script had attempted it, by default the attempt was better irrespective of the quality.'*
>
> *'Initially I reviewed the level of completion of the paper which gave an indication of the engagement with the paper.'*

While most of the abovementioned responses perceived omitted responses as making the process of judging easier, several examiners commented on the difficulty of the judging task arising from omitted responses to some questions.

> *'Blank questions or scripts made it slightly harder to make a judgement.'*
>
> *'I would mentally begin with a general ranking based on questions 4 and 5 and then re-read the lot and check this was accurate throughout the paper. Easier to do when all candidates had answered all questions'*
>
> *'I found this very difficult for this English Language GCSE. I really struggled to decide how to judge, say, a script with no attempt at all on the Reading section but with a perfectly acceptable piece of writing against a script with some weak attempt at all the Reading questions and a flawed, dull piece of writing.'*

While no judges seemed to indicate that they re-marked any scripts, a large number of judges mentioned that they considered the assessment objectives and mark scheme requirements as the basis for their judgements. Thus, their notions of what constitutes good quality in these responses appear to have been strongly based on the mark scheme and the assessment objectives. Some judges indicated that they

'sort of marked', i.e. categorised responses in terms of broad bands or as 'OK', 'good', etc.

> *'The indicative standards were at the front of my mind having just completed marking.'*
>
> *'I would begin reading the response and see how well there were applying the assessments objectives'*
>
> *'I would be mentally marking the scripts. Though I might not assign an actual mark I did find myself placing it into the relevant band'*
>
> *'I ended up allocating some sort of 'mark' (a comment such as, 'OK,' 'good,' or 'rubbish,' for example) for each question.'*

Response features in English language pilots

Some examiners stated that they did not consider specific response features, or allow specific questions, to influence their decision, noting that they focused on overall quality of the responses. However, the majority of examiners identified numerous response features that influenced, and were the basis of, their judgements within the more general approaches discussed above.

There was a range of responses in terms of the level of detail provided. The majority went into considerable detail, suggesting that they used more than one feature to evaluate student responses. Some offered relatively vague responses or no detail at all, although in most cases it could be deduced that they did consider a range of features when making comparisons. The 3 quotes below illustrate these patterns of responses in terms of level of detail:

> *'Quality, clarity and depth of answers when responding to reading tasks. Choice of appropriate detail. With writing tasks – clarity of communication, deployment of vocabulary and linguistic devices, depth and variety of ideas, accuracy of grammar, spelling, punctuation.'*
>
> *'I looked firstly at whether all questions had been attempted. I then looked at the quality of the 10 mark responses, then the writing tasks.'*
>
> *'Just the quality of the responses.'*

The features mentioned can be split into 2 categories: subject-specific features and superficial features. A wide range of subject-specific features was mentioned. Superficial features were considered little. A number of these features, including accuracy (of terminology and understanding), use and understanding of text and source, coherence, response length, incomplete responses, spelling, punctuation and grammar (SPAG), vocabulary and handwriting, were also identified in previous research (Suto and Novakovic, 2012; see also Greatorex, Novakovic and Suto, 2008; Crisp, 2008).

Subject-specific features

The features below are ordered in terms of how frequently they were mentioned in the responses. All of the subject-specific features identified from the responses are present as criteria in English language mark schemes.

Clarity of expression:

> *'The candidates' expression often indicates their own understanding of how language works and so often the better expressed responses also have the higher standard content.'*
>
> *'I looked at clarity of expression,'*
>
> *'Judgement was often based on how precise and/or clear a candidate had been when explaining their ideas'*

Clarity of expression was sometimes mentioned in conjunction with cohesiveness:

> *'Between higher level scripts, the more fluent and cohesive answers demonstrated a higher level of quality.'*
>
> *'I usually compared the openings of each response in terms of which was the most engaging. Then compared the concluding paragraph to look for cohesion in the student's writing.'*
>
> *'Understanding of the question/task and the cohesiveness of their response.'*

Relevance of response to question:

> *'I looked for the detail in which students had responded and if they had answered the question appropriately.'*
>
> *'if students addressed the key skills asked in the questions'*
>
> *'On how the candidate phrased their responses in line with the questions demands'*
>
> *'responses that did not fully answer the question'…. 'reduced the quality of the script'*

Understanding and use of text or source:

> *'To evaluate the overall quality of the scripts, the key criteria was conveying they understood the text.'*
>
> *'Better responses use quotation successfully and support their points with evidence and explanation – in better responses this becomes analysis and so on.'*
>
> *'how clearly does the student understand the writer's intended effect on the reader, how far does the student understand the main ideas in the extract'*

Depth and development of response:

'The depth and development of responses was the critical factor'

'the level of detail and development'

'precision, depth of development of responses'

'depth of explanation/analysis - what were they doing with the bits they'd chosen to discuss'

SPAG and structure:

'accuracy of grammar, spelling, punctuation.'

'Looking at the shape of the text - had they paragraphed their work.'

'the variation in sentence structure, punctuation, the structure of the text'

Vocabulary:

'The last level I use to make a judgement is the type of words that are used. I always give credit when more challenging/mature words are used.'

'The use of vocab and differing sentence forms and overall technical competence.'

'Key vocabulary was sought out in identifying whether the requirements had been met'

Use and identification of devices, for instance, language devices:

'I also looked at which language and structure devices the students had mentioned for question 3; if more sophisticated devices were used successfully then this is normally a better response.'

'knowledge of English Language devices and structures, ability to categorise and manipulate words and phrases'

'on both P1 and P2, I was looking for a mix of language and structural features to be mentioned in the relevant questions.'

Accuracy of terminology and understanding:

'how accurate is the quote/information retrieval in relation to the question (for 1/2 mark questions)'

'Obviously there was accuracy of understanding the reading texts.'

'The accuracy and regularity of subject specific terminology - the more specific the terminology, the better the response (in general)'

Nature of ideas in responses:

'quality/originality/range of ideas.'

'When looking at the 40 mark response I always tend to look at the level of ideas and how they express them. If this is done in a clear way then this response always tends to do better.'

'[…] how ambitious their ideas were.'

Superficial features

Encouragingly, superficial features were commented on least by examiners and seemed to be the least influential in their decision making. Nonetheless, a minority of examiners did identify some features that might be considered superficial.

Response length (usually in conjunction with other features):

'You can often tell just by looking at paragraph length which student has written a more in-depth response.'

'The length of each answer - although some answers were concise and packed with relevant points.'

'I also took into account the length of the responses and the coverage of the text.'

Gut feeling:

'When it came to the writing section, it was more difficult and I just went with my initial reaction.'

'I was encouraged by the instruction to 'go with your gut instinct' and I very deliberately tried not to over-think my judgements.'

Handwriting:

'Some hand writing was difficult to decipher which obviously has an impact. If words are impossible to read it is difficult to be certain whether an answer is correct or not.'

Response features in English literature pilots: examiners vs. teachers

For English literature pilots, this section further demonstrates that both examiners and teachers considered a range of relevant subject specific features in their judgements, most of which were present in the relevant mark schemes. However, the focus here is to explore any differences between teachers and examiners with respect to the features considered. This was done in an attempt to consider possible sources of differences in script rank order that arose between examiner and teacher CJ exercises.

Twelve examiners and 40 teachers responded to the questions: ‘On what basis did you make judgements about relative quality of the scripts in a pair?’, and ‘Were there any particular script features and/or responses to particular questions which influenced your judgements?’ A thematic analysis was performed using the QSR International's NVivo 11 software (2015), wherein the types of influential features identified were compared between experts and teachers.

It should be noted that, given very different sample sizes, it is likely that some features that appear to only be present in teacher responses may have appeared in examiner responses in a bigger sample of examiners. Nevertheless it may be informative to investigate the prevalence of different response features in the 2 samples as well as any discrepancies between them.

In terms of aspects of the general approach to judging, similarly to the English language pilots, both teachers and examiners considered assessment objectives, challenge of text or question and sometimes focused on specific questions or specific parts of the response. Some examiners also noted that missing responses influenced their judgements, sometimes making them easier and sometimes more difficult.

In terms of the response features, the table below summarises those that were noted by examiners and/or teachers as contributing to their judgements of script quality. The table also shows the prevalence of these features in responses and whether the features are present in the mark scheme or not. The features are ordered in descending order of prevalence in examiner responses.

Several observations can be made with respect to the patterns apparent in the data. The majority of the features mentioned appear to be clearly subject-specific, although there were several features which can be seen as superficial, such as handwriting or length of response. Other features such as SPAG, independent thoughts or textual comparisons were mentioned by a few participants, but are not explicitly rewarded by the mark schemes. This could suggest that they are not explicitly part of the construct assessed in these examinations, although they may be otherwise considered as valid differentiating features that may support holistic qualitative judgements.

Examiners tended to mention fewer influential features (4 at most) in their answers compared to teachers who were more likely to list numerous features. However, overall, the majority of the features were considered by both groups at least to some extent. The top 9 features considered by the examiners were also among most frequently considered by the teachers. All of these 9 features are also present in the mark scheme and appear to constitute the key aspects of the construct assessed in English literature.

Table 54 *Response features considered by examiners and teachers in English literature pilots*

Response feature	N mentions (examiners)	N mentions (teachers)	In mark scheme?
Coherence	9	28	Y
Understanding and use of the text	8	35	Y
Relevance of response to the text	6	14	Y
Structure	5	7	Y
Argument	4	22	Y
Context	4	14	Y
General content	4	13	Y
Relevance	4	6	Y
Analysis	3	25	Y
Textual comparison	2	2	N
Length	2	1	N
Consideration of the author	1	6	Y
Handwriting	1	5	N
Terminology	1	13	Y
Use of linguistic devices	1	6	Y
SPAG	1	5	N
Depth of analysis/understanding	0	10	Y
Accuracy	0	8	Y
Reference to another source	0	5	Y
Independent thoughts	0	3	N

In addition to these key features, a number of other features, also present in the mark scheme, were considered mostly or only by the teachers. Among these, terminology, depth of analysis/understanding, accuracy, consideration of the author and use of linguistic devices were the most frequent. While the more detailed features considered by the teachers perhaps suggest a somewhat different approach to judging (for example, paying more attention to detail in responses while judging), the fact that these features appear in the relevant mark schemes and are relevant to the construct of these assessments suggests that the judgements based on them would not have been less valid than those of the examiners.

However, some of the abovementioned superficial features or those not present in the mark scheme were more prevalent amongst teachers (e.g. handwriting, SPAG, independent thoughts), and their consideration could have led to candidates' work being evaluated differently compared to how it would have been evaluated using the mark scheme. It is, however, reassuring that only a small number of features, mentioned by relatively few participants, belong in this category.

Discussion

Summary and key findings

The results of both the RO and the PCJ larger-scale pilots in GCSE English language consistently show that the comparative judgement exercises succeeded in producing plausible script quality scales (with SSRs of 0.9 or higher) and high level of agreement between original test score scales and quality measure scales (correlations of 0.9 or higher). Furthermore, the smaller-scale pilots in GCSE English language, which collected around 20 judgements per script and were based on about 50% of the mark range, appeared to work equally well as the larger-scale pilots with 25 comparisons per script and 70% of the mark range. This may suggest some scope for streamlining data collection in operational settings.

In the PCJ pilots in subjects other than GCSE English language, conducted with only 10-12 comparisons per script, the SSRs ranged from 0.7 to 0.8. Mark-measure correlations tended to vary between 0.6 and 0.8 in these pilots. While most of the grade boundary estimates from these smaller pilots were still plausible, we suggest caution in a few cases where either mark-measure correlations or the SSRs, alongside very wide confidence intervals, resulted from the pilots. Except for some English literature pilots, the RO pilots conducted in subjects other than GCSE English language, where 20-36 comparisons were collected per script, all produced SSRs of 0.9 or higher, mark-measure correlations of 0.75 or higher and largely plausible grade boundary estimates.

Some English literature pilots in particular in some cases resulted in fairly low mark-measure correlations and/or SSRs lower than 0.7. This was particularly prominent in the RO pilots in English literature 1 P2 and English literature 2 P1, as well as in the English literature 2 teacher PCJ exercise. The average of 20 comparisons per script did not seem to help boost mark-measure correlations despite reasonable SSRs in these RO exercises, while both SSRs and correlations were less than optimal in teacher PCJ exercises, based on 12 comparisons per script.

As we noted previously, it is possible that there were other factors at play here that weakened the mark-measure correlations for English literature specifications. For instance, marking reliability of original markers may have been low to begin with, leading to a different order of quality compared to the one that resulted from larger-scale judging exercise. In fact, we do know from other research that marking reliability for English literature is relatively low compared to some other subjects. However, this is also true of English language, yet the mark-measure correlations in these pilots were all very high.

Another difference between the English literature specification and English language specifications is that the mark scales for English literature where the correlations were particularly low were shorter than those for English language, and thus possibly less discriminating to begin with. Therefore, it may have also been more difficult for English literature judges to discriminate between scripts when judging holistically, leading to more variability in judgements and less agreement with the original test score scale, especially given the smaller number of judgements collected in some of these exercises.

Another possibility is that, in some cases, low mark-measure correlations could be a sign of incongruence between the constructs that the judges considered important when making holistic judgements vs. those that were rewarded by the relevant mark schemes. We found some evidence of these incongruences based on the analysis of the features the examiners and teachers reported as influential in their holistic judgements. This could to some extent explain relatively low correlation between the measures produced by teacher judgements vs. those of the examiners for the same specification in the English literature 2 pilots. However, there was no clear evidence of incongruence with the mark scheme in the responses of the examiners themselves that would explain the low correlations in examiner RO exercises for English literature 2. Therefore, a lot of the variability in the outcomes, and low correlations, may be mainly attributable to the fact that relatively small number of judgements was collected in these pilots in the first place, alongside the short test score scales leading to low discriminability of scripts and necessarily incongruences with the original mark rank order.

It would, however, seem important to investigate possible incongruence between what is considered important when judging holistically compared to what is rewarded in mark schemes. Congruence with the mark scheme could be seen as an important aspect of validity of CJ exercises, especially if the judgmental methods are to be used for standard maintaining. On the other hand, these kinds of holistic judging exercises could also point to issues with mark schemes themselves, if it became apparent that they reward aspects of performance that may not provide the best evidence of the constructs considered important beyond the constraints of individual assessment instruments, which could emerge in holistic judging.

When we trialled the pinpointing approach, this failed to result in convincing script quality scales, and produced some implausible grade boundary estimates. We would suggest that this approach, with its focus on a narrow range of mark points around the grade boundaries, may not be the optimal way of maximising judgement consistency and scale reliability.

Reassuringly, the results of the pilots which were carried out on the same specifications largely cross-validated each other even where smaller number of comparisons per script were collected, producing very similar grade boundaries, while the script quality measures from these pilots were mostly highly correlated. All this suggests that pooling sufficiently large number of judgements over most of the effective test score scale can increase the validity of the outcome of expert judgement, and, thus our confidence in expert judgement recommendations.

Further evaluation and considerations

Grade boundary outcomes based on rank ordering and paired comparisons for individual specifications were mostly credible, with some exceptions in specifications where lower SSRs or mark-measure correlations were achieved. These sub-optimal scale properties, alongside wide bootstrapping confidence intervals, suggest caution with respect to taking some of the boundary estimates at face value. While in most cases, good statistical evaluation results were associated with grade boundaries that were entirely congruent with the Y2 operational ones, in some cases, they were associated with the boundaries that were fairly discrepant from the Y2 operational boundaries. In the latter case, in particular, it would be necessary to consider a

range of available sources of evidence and give appropriate weight to these sources in deciding on the most likely correct grade boundaries.

English language is a good example of the latter pattern, where both paper and qualification level pilot boundaries tended to be lower than the Y2 operational ones. The apparent pattern of lower pilot grade boundaries also suggested a possibility that the results might be an artefact of the way pilots were implemented. For instance, the judges were in all cases aware which scripts came from which session, and may have used this to direct their judgements in some way. However, examining the patterns of differences between pilot and operational boundaries across all the pilots suggests that there is little evidence of consistent positive or negative differences (where, for instance, consistently lower pilot boundaries would have led to higher outcomes for candidates). Thus, it seems unlikely that the results were consistently affected either by idiosyncrasies of the rank ordering or the paired comparisons methods, or by deliberate “gaming” by the judges. Furthermore, the comparison of consistency levels in the within- vs. between-session judgements did not reveal any worrying differences that would suggest the between-session comparisons may have been less consistent or degrading the measurement process.

Considering the bootstrapping results, it is clear that there is more potential variability for GCSE at grade 1 in particular, but also grade 9, and at AS level to some extent at grade E. This is a familiar effect with respect to extreme scores. It would be important to consider how to overcome this challenge in judgemental exercises, if we are to have sufficient confidence in their outcomes even for extreme scores.

Regarding using the confidence intervals derived from bootstrapping to quantify likely variability in comparative judgement outcomes, we have argued that traditional confidence intervals based on +/-2SD around the mean might be too stringent in this context. We suggested that middle 50% IQRs might be appropriate, especially within the constraints of exercises with similar design parameters and judge expertise year on year. This may also be considered appropriate given the apparent robustness of these methods to a range of design and other manipulations, as well as replicability of the results in different contexts and with different judges.

With respect to judges’ ability to compensate for differences in paper difficulty between sessions in their script quality judgements, which is one of the assumptions of comparative judgemental methods when used for standard maintaining, there is some indication that the judges may be able to do this to some extent. They referred to reasonable techniques of doing so when accounting for this in their survey responses. However, it is not easy to see very clear patterns of alignment between judges’ initial views of paper difficulty and the corresponding pilot outcomes in most cases. Furthermore, a large number of judges thought that the papers from different sessions were similar, often irrespective of the final outcome. Arguably, however, most of the grade boundary differences between sessions were indeed very small, and possibly justify a view that papers where boundaries between sessions differ by 1 or 2 marks can reasonably be described as similar. Additionally, for some of the papers in this pilot, some of the approaches to marking might change between years, for instance in terms of changes in leniency or severity to awarding top level mark bands. This means that there is not a straightforward link between strict paper difficulty (i.e., paper difficulty as an aggregate of the tasks, not the marking

approach) and the relationship between the 2 mark scales. All of this is an area that needs more exploration.

It would, however, seem important to be realistic about the level to which judges can reasonably be expected to be able to account for differences in empirical test difficulty to the extent a statistical equating method based on large quantities of data could. Such expectations could lead to setting unrealistic evaluation targets for a judgemental method that is essentially a proxy for statistical equating in the absence of a more appropriate method. This probably needs to be recognised as an unavoidable source of error and a shortcoming of all approaches to standard maintaining that do not rely on pre-testing test items routinely to enable robust statistical equating methods to be used.

Our qualitative analysis of the strategies and response features that were reported by the judges as influential in their holistic judgements for English language and English literature suggests that the judges used mostly valid strategies and response features when making their judgements, which accord with those identified in prior research and were mostly present in relevant mark schemes. However, it should be borne in mind that these were reported post-hoc in surveys and may not provide sufficiently in-depth, coherent or detailed picture of the judging process in general or for individuals.

Some pilots were designed in such a way as to facilitate achieving SSRs of around 0.9. However, lower levels of reliability might be considered appropriate in certain contexts. Our analyses on sections of our data with smaller number of comparisons per script suggest that there may be scope to reduce the scale of these exercises somewhat, but that going below 10 comparisons per script might lead to too much variability in the outcomes, reducing our confidence in the results.

With respect to optimal number of mark points and scripts to include in CJ exercises (for instance, in our case, 50% vs. 70% of mark points), it should be noted that, ideally, the sample of scripts used in CJ exercises should be in some way representative of the full set of scripts from the relevant examination (cf. Benton, 2019). Reducing the number of mark points included in CJ exercises would reduce the representativeness of the sample further, potentially leading to grade boundary outcomes that would not be representative of the outcomes that would have been obtained if the full set of mark points and scripts was judged. Furthermore, where smaller number of scripts is used, it would be important to consider the implications this would have for bootstrapping analysis, as its results may be less valid (for example, may appear to overestimate the variability in the outcomes) when there is a small number of objects in the sampling pool. While the number of scripts to be included in CJ exercises may be limited by practical considerations in operational contexts, the precise impact of different sample sizes and profiles requires further research.

Regarding the effect of including or removing imputed measures from grade boundary estimation, it would appear that this can have at least as much impact as reducing the number of comparisons per script. It would seem important to investigate these effects further and ensure that decisions about whether to include or exclude imputed measures are not arbitrary and are documented in reporting the results.

We also looked at the effects on judge expertise (ordinary vs. senior examiner) on CJ outcomes, concluding that there does not appear to be a tangible and consistent effect of this. This suggests, alongside other research (e.g. Verhavert et al., ibid.; Raikes, Scorey and Shiell, ibid.) that ordinary examiners can participate in these exercises without compromising our confidence in the outcomes. This could potentially help organise CJ exercises during live marking, as it could free up senior examiners to deal with their other obligations during this period.

These pilots suggest a range of options for running CJ exercises on different scales and give an indication of the likely robustness of the results based on them. We have also pointed out a number of advantages of the CJ methods for capturing expert judgement, as well as considerations that need to be taken into account when planning and carrying out these exercises. Below we highlight some operational implications of including these methods in awarding as well as some of the as yet unanswered questions for further research.

Operational implications

There are a number of operational implications and considerations when understanding what these judgemental processes might look like within the context of awarding. This section of the discussion attempts to give some high level analysis of what these implications and considerations might be in terms of how comparative judgemental methods might feasibly work effectively and efficiently: how the outputs might work within a live awarding context, and some consideration of implementation costs (both the one-off implementation costs, as well as the 'business as usual' costs). A consideration of these potential operational implications would likely form an important aspect for any deliberations around the potential for whether and how the current process might change.

As described previously, none of these pilots took place fully in the context of an award. Including them fully into the awarding process would entail both of two key aspects:

1. the judgements and the analysis being conducted prior to the actual award
2. the outcomes of the analysis of this judgemental exercise feeding into the awarding meeting decisions

In relation to the first of these aspects, some of the pilots[22] were conducted prior to the actual award, i.e. after or towards the end of marking and prior to awarding. Because of this, we can reflect upon some of the likely operational implications for comparative judgement or rank ordering methods as a judgemental activity at this point. In relation to the second of these aspects, we can highlight some potential practical benefits as well as challenges.

Operational implications for using comparative judgement methods routinely

Timing

Unlike the current method of capturing expert judgement, one potential advantage of this method is that it does not have to take place after a high proportion of the

[22] Media studies and English language

marking has been complete. Because the range of scripts involved does not depend upon knowing the full mark distribution or modelling the qualification outcomes, but rather is sampled across the (main part of) the mark range, then scripts can be selected from the current year when around 40-50% of the marks are on the system for each component, rather than when around 85% of the marks are on the system across all components. This should afford a more flexible window for conducting the expert judgemental exercise.

Feedback from the awarding organisation and the judges involved in live CJ pilots suggested that the pilots did not impact negatively on completion of operational marking and other operational processes during the period leading up to the awarding meeting. The survey results for English language did indicate that more senior examiners occasionally struggled with workload that participation in the pilots created at the time when they would normally still be marking or reviewing other markers' work. The ordinary examiners amongst our judges did not comment on any negative impact this had on their marking as this had usually already been completed prior to taking part in the pilots.

Preparation and distribution of materials

Preparation of the materials is an important aspect to consider in the operationalisation of these methods. In the current awarding process, relatively fewer scripts are used. For example, around 5-15 scripts might be identified per mark point to be considered for any key grade, with extras available should the range need to be increased. There might be some criteria applied too – such as only those scripts marked by senior examiners, or higher grade examiners. Over the years, exam boards have developed automated systems to identify the scripts on marks in the identified range. And, increasingly, where the judgemental aspect of awarding takes place remotely, electronic versions of scripts are distributed through specific software, removing the need for obtaining a physical copy of a script from storage.

With potentially a new system of capturing expert judgement, different software would most likely need to be developed in order to select and distribute scripts:

- For paired comparative judgement, we used an existing electronic platform which requires pdf versions of scripts to be uploaded, once they have been manually selected. For a smoother process, ideally such a system would need to be more integrated into the systems which store scripts to enable greater opportunities for automation. We are aware of systems which have such capability, including for portfolio-based assessments (such as Kimbell, 2011).
- For rank ordering, the paper-based methods used in the context of these pilots were time-consuming and required manually sorting scripts into packs and envelopes. While we used the same template of pack design (only one is needed for each maximum mark range), in order to work in a real life context there would have to be greater automation in terms of collating and distributing the packs. For paper-based approaches, this could take the form of an 'intelligent printing' system whereby scripts according to their pack design can be automatically printed, collated and dispatched to judges; or indeed a fully electronic system might be possible. There have been systems we are aware of which facilitate the distribution of pairs of work and collect the judgements (such as that reported in Kimbell, ibid) – it is possible these could

be extended to rank ordering. There are others currently used as well as in development.

Judges

The CJ methods allow for greater flexibility in terms of judges compared to the current processes. While the current process requires senior examiners (usually principal examiners, lead markers or lead setters), this can be widened to a different population of experts, including less senior examiners and/or teachers in CJ exercises. In one of the pilots we conducted, for example, 40 teachers made a small number of judgements, rather than a small number of judges making many judgements. Again, this can add to the flexibility of this method when it is conducted. Part of the flexibility in the judges is that the nature of the intuitive task requires no specific training beyond a certain level of subject expertise.

Duration and cost of capturing expert judgement

The pilots involved different models in respect of the numbers of judgements per script and the number of scripts. For those models with lower numbers of judgements and scripts (around 10 judgements per script), for paired CJ in particular, the amount of contracted time for judges ('judge days') was probably very similar to the current method of capturing expert judgement. However, while this was probably sufficient, higher numbers of judgements per script are preferable and this would necessarily entail an increased number of judge days. However, given that this method of capturing expert judgement has a number of key advantages, it is perhaps not surprising that it is likely to require some additional resource.

Table 55: *Comparison of the current method of capturing expert judgement compared to comparative methods in terms of operational and implementation considerations*

	Current method	**Comparative Judgement methods**
Timing	After the completion of marking, before awarding.	After the completion of ≈40% of marking, before awarding.
Preparation and distribution	Mainly remote, bespoke software systems.	Some investment and development needed to create software and systems to facilitate.
Location	Face to face or remote	Face to face or remote.
Medium	Paper based or electronic.	Paper based or electronic. NB some development work likely required.
Judges	Generally, 4 to 6 judges per qualification; usually senior examiners.	Probably no less than 6 judges, but with greater flexibility in terms of seniority and how the desired number of judgements per scripts can be distributed. For example, it could be a smaller number of judgements across a broader range of examiners and/or teachers, providing that they have sufficient subject knowledge.
Training of judges	There is usually some training of judges at least for the first award. They also need to be familiar with	Little or no training is required as the judgement (comparison of scripts) is essentially an intuitive

	the archive scripts so that they can make appropriate judgements about the current year's scripts' gradeworthiness.	task, providing that a judge has the appropriate expertise.
Duration and cost of judges	Estimated around 6-16 judge days per award.	Similar for paired CJ with fewer (≈10) judgements per script. For a greater number of judgements per script, the number of judge days could double.
Analyses required and interpretation	Simple: the patterns of ticks and crosses for each mark point can be 'eye-balled'.	More sophisticated analyses are required. Some aspects of these can be automated within the software used for capturing expert judgement (e.g. fitting the Rasch model and deriving a measure of script quality). However, some aspects may need further automation (regressing mark on measure) or require expert human judgement (inspecting model or judge fit, considering mark measure correlations).

Consideration of operational implications for using CJ outcomes to feed into the awarding meeting decisions

As described previously, none of the pilots took place fully in the context of an award in that the outcomes of the analyses did not feed into the decision making in the awarding process. As such, there are no 'lessons learned'. However, we can highlight some potential practical benefits as well as challenges in relation to how this might work in practice.

Once the analyses have taken place, this form of expert judgement should identify an equivalent mark for the same performance standards across the 2 years in question for each key grade boundary. Where this equivalent mark is the same or very close (within one mark) of the statistically recommended boundary, this might mean that the final decision of where to place the boundary is very straightforward.

Where the statistics and the judgement point to different places, there is likely to be more to think about. In effect, the chair of examiners, or the persons charged with the grade boundary recommendations, will likely need to carefully weigh the evidence from these 2 different sources in order to establish the relative credibility. This might mean a series of questions about each of the sources of evidence. For example, to establish the credibility of the judgemental exercise, the chair of examiners might want to understand the various quality indicators of the data as well as evaluate its overall plausibility. This might include, for example:

- whether the SSRs and the mark-measure correlations are sufficiently high
- the profile of judge and script fit
- the extent to which the indicative boundaries are similar or different from previous years

In relation to the statistically recommended boundaries, they awarding panel would also be reviewing some features of the predictions including:

- the stability of the size and nature of the cohort between the years
- the proportion of matched candidates

There are some different methods to integrate the judgemental exercise into awarding, depending upon, prima facie, the relative weight which might be accorded to the two main sources of evidence. This relative balance of the 2 sources of evidence is, naturally, on a continuum. To help consider this continuum, the following represent different ways of regarding relative balance of the 2 sources of evidence:

1. assume the statistical predictions carry the greatest weight. The judgementally derived boundaries should only suggest divergence from the statistical boundaries if they indicate a significant difference[23] from the statistically recommended boundary and the various quality indicators are strong and/or have been consistent over a number of sessions or series
2. assume the statistical predictions carry greater weight. The judgementally derived boundaries might suggest some modification to statistically derived boundaries where either there is some degree of uncertainty in some of the assumptions necessary for the statistics and the judgemental exercise has met all of the quality indicators
3. assume equal weight between the statistical predictions and the judgemental source of evidence, and interrogate both sources of evidence for credibility
4. assume the judgemental exercise carries the greatest weight

One implication, however, is that the awarding panel and/or chair would need advising or training on how to appraise these different sources of evidence and weigh accordingly.

Further work

Further research into some of the more technical issues that might have some impact on the results will be undertaken. This could include reanalysing the rank ordering data using the rank ordering model of analysis rather than the paired comparisons model to estimate the level of possible SSR overestimation in RO exercises. It may also be useful to investigate the extent of scale inflation due to model overfit as well as the effect of different methods of measure imputation on scale properties. Another possibly useful avenue would be to explore using measure on mark rather than mark on measure regression to estimate Y2 grade boundaries.

Conclusions

Even though further consideration needs to be given to the merits of different CJ methods and specific designs in operational contexts, overall, the results of our pilots suggest that CJ methods are very promising for capturing expert judgement for the purpose of standard maintaining. The totality of the pilots indicate that pooling a sufficiently large number of judgements over most of the effective test score scale can increase the reliability of the outcome of expert judgement, potentially increase the validity of expert judgement in standard maintaining, and thus increase our confidence in expert judgement recommendations. The fact that CJ methods are

[23] For example, the statistically recommended boundary is not within the 95% confidence interval indicated by the bootstrapping.

implemented independently of statistical grade boundary recommendations and knowledge of original script marks helps to keep judgemental evidence as an independent source of evidence that could be attributed its own weight appropriate to the specific context of use.

Some of our pilots were designed in such a way as to facilitate achieving SSRs of around 0.9. However, lower levels of reliability might be considered appropriate, and therefore smaller-scale exercises, which might still be sufficiently robust, may be reasonably attempted in some contexts. Decisions about the scale of CJ exercises might also need to be driven by the intended weight that might be given to judgemental evidence in each case. Where more weight might be placed on the judgemental outcomes (for instance, where there is less confidence in the statistical outcomes for whatever reason) it might be reasonable to collect judgemental data on a larger scale.

It would also seem important to continue investigating suitability of different criteria for evaluating the comparative judgement methods, including appropriate confidence intervals for grade boundary estimate precision. Evaluation criteria for judgemental methods might to some extent depend on the way we conceptualise their place in awarding. While these methods certainly go a long way towards enhancing the reliability of expert judgement and increasing our confidence in its recommendations, it may still be inappropriate to attempt to evaluate them according to stringent criteria that may be applicable for purely statistical methods of equating.

References

Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). Estimating item difficulty with comparative judgements. *ETS Research Report Series*, 2014, 1–8. Retrieved from https://doi.org/10.1002/ets2.12042

Baird, J. (2000). Are Examination Standards all in the Head? Experiments with Examiners' Judgements of Standards in A Level Examinations. *Research in Education,* 64, 91–100. Retrieved from https://doi.org/10.7227/RIE.64.9

Baird, J. & Dhillon, D. (2005*). Qualitative Expert Judgements on Examination Standards: Valid, but Inexact*. AQA research report RPA_05_JB_RP_077. Guildford: AQA.

Benton, T. (2019). *Maintaining standards using paired comparisons (but without Bradley or Terry)*. Unpublished internal report. Cambridge: Cambridge Assessment.

Benton, T. & Bramley. T. (2015). *The use of evidence in setting and maintaining standards in GCSEs and A levels*. Discussion paper. Cambridge: Cambridge Assessment. Retrieved from https://www.cambridgeassessment.org.uk/Images/459318-the-use-of-evidence-in-setting-and-maintaining-standards-in-gcses-and-a-levels.pdf

Benton, T. & Elliott, G. (2016). The reliability of setting grade boundaries using comparative judgement, *Research Papers in Education*, 31(3), 352–376. Retrieved from http://dx.doi.org/10.1080/02671522.2015.1027723

Benton, T. & Gallacher, T. (2018). Is comparative judgement just a quick form of multiple marking? *Research Matters*, 26, 22–28. Retrieved from https://www.cambridgeassessment.org.uk/Images/514231-research-matters-26-autumn-2018.pdf

Black, B. & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education*, 23(3), 357–373. Retrieved from http://dx.doi.org/10.1080/02671520701755440

Black, B. (2008). *Using an adapted rank-ordering method to investigate January versus June awarding standards*. A paper presented at the Fourth Biennial EARLI/Northumbria Assessment Conference, Berlin, Germany, August 2008. Retrieved from https://www.cambridgeassessment.org.uk/Images/109767-using-an-adapted-rank-ordering-method-to-investigate-january-versus-june-awarding-standards.pdf

Black, B. & Gill, T. (2008). *Using Rank-Order as a method for standard maintaining in small entry units*. Unpublished internal report. Cambridge: Oxford Cambridge and RSA Examinations.

Bradley R.A. & Terry M.E. (1952). Rank Analysis of Incomplete Block Designs I: The Method of Paired Comparisons. *Biometrika*, *39*, 324–45. Retrieved from http://dx.doi.org/10.2307/2334029

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement 6(2),* 202–23. Retrieved from https://www.researchgate.net/publication/7941436_A_rank-ordering_method_for_equating_tests_by_expert_judgment

Bramley, T. (2007). Paired comparison methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246–300). London, UK: Qualifications and Curriculum Authority. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/487059/2007-comparability-exam-standards-i-chapter7.pdf

Bramley, T. (2012). The effect of manipulating features of examinees' scripts on their perceived quality. *Research Matters: A Cambridge Assessment Publication, 13*, 18-26. Retrieved from https://www.cambridgeassessment.org.uk/Images/469827-the-effect-of-manipulating-features-of-examinees-scripts-on-their-perceived-quality.pdf

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

Bramley, T. & Gill, T. (2010). Evaluating the rank-ordering method for standard maintaining, *Research Papers in Education*, 25(3), 293–317. Retrieved from http://dx.doi.org/10.1080/02671522.2010.498147

Bramley, T. & Vidal Rodeiro, C.L. (2014). *Using statistical equating for standard maintaining in GCSEs and A levels.* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Retrieved from https://www.cambridgeassessment.org.uk/Images/182461-using-statistical-equating-for-standard-maintaining-in-gcses-and-a-levels.pdf

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. http://dx.org/10.1080/0969594X.2017.1418734

Brandon, P. R. (2004). Conclusions About Frequently Studied Modified Angoff Standard-Setting Topics. *Applied Measurement in Education*, 17(1), 59-88.

Chambers, L., Vitello, S., & Vidal Rodeiro, C.L. (2019, November). *Moderation of non-exam assessments: a novel approach using comparative judgement.* Paper presented at the 20th annual AEA-Europe conference, Lisbon, Portugal.

Christensen, K. B., Kreiner, S., & Mesbah, M. (2013). *Rasch Models in Health*. London, UK: Wiley & Sons.

Cresswell, M. J. (1997). *Examining Judgements: Theory and Practice of Awarding Examination Grades*. Unpublished PhD thesis, University of London Institute of Education, London.

Crisp, B. (2008). Do assessors pay attention to appropriate features of student work when making assessment judgements? *Research Matters: A Cambridge Assessment publication*, 6, 5-9. Retrieved from https://www.cambridgeassessment.org.uk/Images/109983-research-matters-06-june-2008.pdf

Cuff, B., Meadows, M. & Black, B. (2018). An investigation into the Sawtooth Effect in secondary school assessments in England. *Assessment in Education: Principles, Policy & Practice*, 26(3), 321–339. Retrieved from https://doi.org/10.1080/0969594X.2018.1513907

Curcin, M., Black, B., & Bramley, T. (2010). *Towards a suitable method for standard-maintaining in multiple-choice tests: capturing expert judgment of test*

difficulty through rank-ordering. Paper presented at the Association for Educational Assessment-Europe (AEA-Europe) annual conference, Oslo, Norway.

Elliott, G., & Greatorex. J. (2002). A fair comparison? The evolution of methods of comparability in national assessment. *Educational Studies 28(*3), 253–264. Retrieved from http://dx.doi.org/10.1080/0305569022000003670

Forster, M. (2005). *Can Examiners Successfully Distinguish Between Scripts that Vary by only a Small Range on Marks?* Unpublished internal paper. Cambridge: Oxford Cambridge and RSA Examinations.

Gill, T., & Bramley, T. (2013). How Accurate are Examiners' Holistic Judgements of Script Quality? *Assessment in Education: Principles, Policy and Practice,* 20(3), 1–17. Retrieved from http://dx.doi.org/10.1080/0969594X.2013.779229

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.

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

Gray, E. (2000). *A comparability study in GCSE science 1998. A study based on the summer 1998 examination*. Organised by Oxford Cambridge and RSA Examinations (Midland Examining Group) on behalf of the Joint Forum for the GCSE and GCE.

Greatorex, Novakovic and Suto, 2008;

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

Holmes, S. D., Meadows, M., Stockford, I. & He, Q. (2018). Investigating the Comparability of Examination Difficulty Using Comparative Judgement and Rasch Modelling. *International Journal of Testing*, 18(4), 366-391. Retrieved from https://doi.org/10.1080/15305058.2018.1486316

Impara, J.C., & Plake, B.S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions of the Angoff standard setting method. *Journal of Educational Measurement*, 35, 69–81.

Jones, B., Meadows, M., & Al-Bayatti, M. (2004). *Report of the inter-awarding body comparability study of GCSE religious studies (full course) summer 2003*. Assessment and Qualifications Alliance.

Jones, I., Swan, M., & Pollitt, A. (2014). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, 13, 151–177. Retrieved from https://doi.org/10.1007/s10763-013-9497-6

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

Jones, I., Wheadon, C., Humphries, S. & Inglis, M. (2016). Fifty years of A-level mathematics: have standards changed? British Educational Research Journal, 42 (4), 543–560. Retrieved from https://doi.org/10.1002/berj.3224

Kimbell, R., Wheeler, T., Stables, K., Shepard, T., Martin, F., Davies, D., … Whitehouse, G. (2009). *E-scape portfolio assessment phase 3 report*. London: Goldsmiths, University of London. Retrieved from https://www.teachertoolkit.co.uk/wp-content/uploads/2014/08/e-scape_phase3_report.pdf

Kimbell, R. (2011). Evolving project e-scape for national assessment. *International Journal of Technology and Design Education*, 22(2), 135–155. Retrieved from https://link.springer.com/article/10.1007/s10798-011-9190-4

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices (2nd ed.).* New York: Springer.

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

Linacre, J. M. (1992). Objective measurement of rank-ordered objects. In Mark Wilson (Ed.) *Objective measurement: Theory into practice, Volume 1*. Norwood, NJ: Ablex.

Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions*, 16:2, 878. Retrieved from https://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2011). *A user's guide to FACETS Rasch-model computer programs*. Program Manual 3.68.1. Chicago, IL: Winsteps.com.

Linacre, J. M. (2019). *Facets computer program for many-facet Rasch measurement*, version 3.81.2. Chicago, IL: Winsteps.com.

Newton, P. E. (2011). A level pass rates and the enduring myth of norm-referencing. *Research Matters: A Cambridge Assessment Publication, Special Issue 2: comparability*, 20–26. Retrieved from https://www.cambridgeassessment.org.uk/Images/109991-research-matters-special-issue-2-comparability.pdf

Nunnally, J. C. (1978). *Psychometric theory (2nd ed.).* New York, NY: McGraw-Hill.

Ofqual (2019a). Summer 2019 Data Exchange Procedures. GCE (AS/A level), GCSE and Project Qualifications. Coventry, UK: Ofqual.

Ofqual (2019b). Summer 2019 GCSE, AS, A level and level 3 project qualifications. A summary of our monitoring. Coventry, UK: Ofqual. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/826699/Summer_2019_monitoring_summary_-_FINAL196533.pdf

Pollitt, A. (2004, June). *Let's stop marking exams*. Paper presented at the annual conference of the International Association for Educational Assessment, Philadelphia, USA.

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

Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22: 157–170. Retrieved from http://dx.doi.org/10.1007/s10798-011-9189-x

Pollitt, A., & Elliott, G. (2003). *Monitoring and investigating comparability: A proper role for human judgement*. Cambridge: Research and Evaluation Division, University of Cambridge Local Examinations Syndicate. Retrieved from https://www.researchgate.net/publication/251218769_Monitoring_and_investigating_comparability_a_proper_role_for_human_judgement

Raikes, N., Scorey, S. & Shiell, H. (2008, September). *Grading Examinations using Expert Judgements from a Diverse Pool of Judges*. Paper presented to the 34th annual conference of the International Association for Educational Assessment, Cambridge.

Rhead, S., Black, B. & Pinot de Moira, A. (2018). *Marking consistency metrics: An update.* (Report No. Ofqual/18/6449/2). Coventry, UK: Ofqual. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/759207/Marking_consistency_metrics_-_an_update_-_FINAL64492.pdf

Robitzsch, A. (2019). *Sirt: Supplementary item response theory models*. R package version 3.7–40. Retrieved from https://cran.r-project.org/web/packages/sirt/index.html

Scharaschkin, A. & Baird, J. (2000). The Effects of Consistency of Performance on A Level Examiners' Judgements of Standards. *British Educational Research Journal* 26, 343–357. Retrieved from https://doi.org/10.1080/713651557

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

Stringer, N. (2012). Setting and maintaining GCSE and GCE grading standards: the case for contextualised cohort-referencing. *Research Papers in Education*, 27(5), 535–554. Retrieved from https://doi.org/10.1080/02671522.2011.580364

Suto, I. & Novakovic, N. (2012). An exploration of the examination script features that most influence expert judgements in three methods of evaluating script quality. *Assessment in Education: Principles, Policy & Practice*, 19(3), 301-320. Retrieved from https://doi.org/10.1080/0969594X.2011.592971

Taylor, R. & Opposs, D. (2018) 'Standard setting in England: A levels'. In Baird, J., Isaacs, T., Opposs, D. & Gray, L. (eds.) *Examination standards: how measures & meanings differ around the world*. London: UCL IOE Press.

Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review* 3: 273–86. Retrieved from https://doi.org/10.1037/h0070288

Thurstone, L.L. (1931). Rank order as a psychophysical method. *Journal of Experimental Psychology* 14: 187–201. Retrieved from https://psycnet.apa.org/doi/10.1037/h0070025

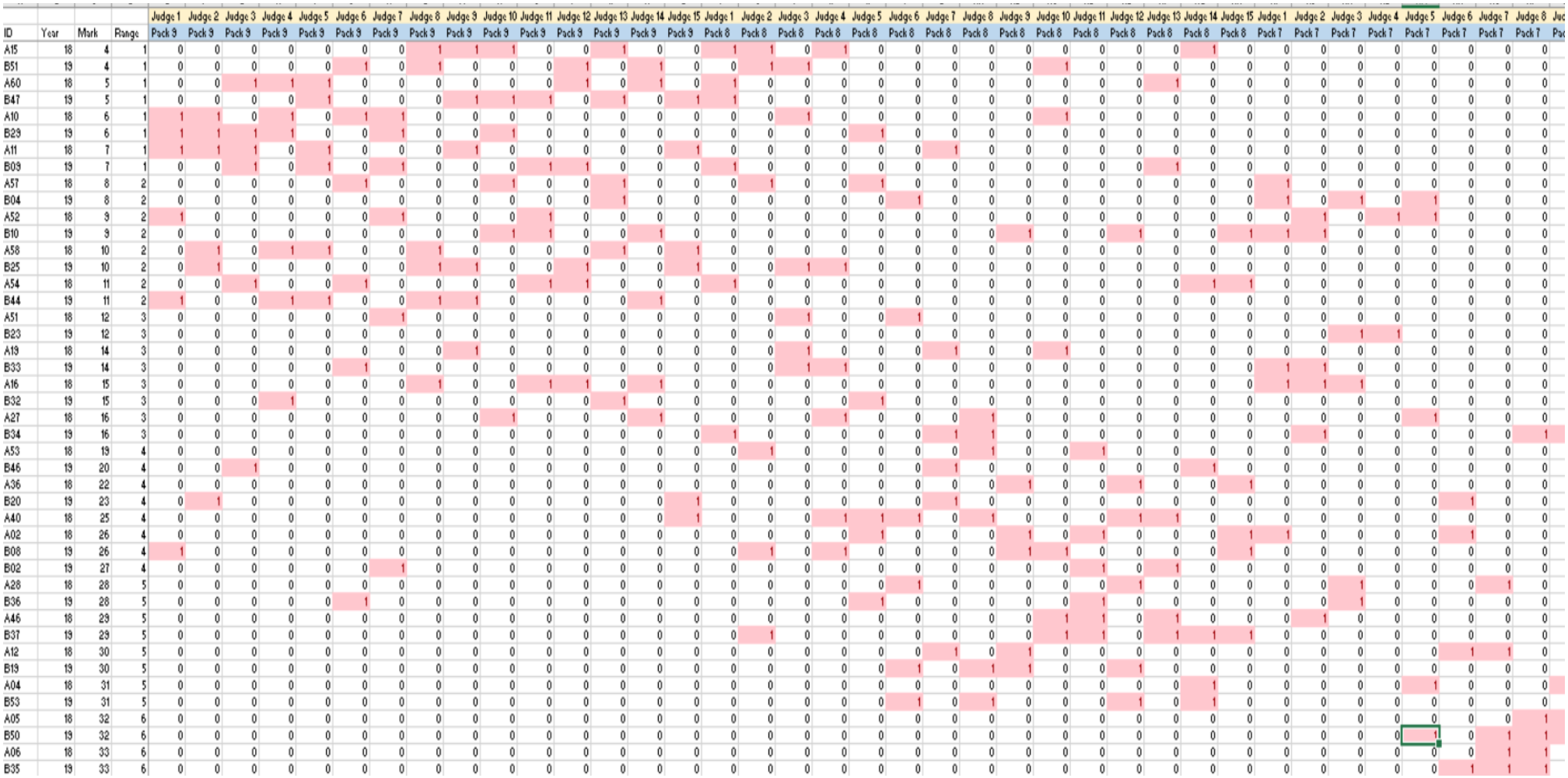
Verhavert, S., De Maeyer, S., Donche, V. & Coertjens, L. (2018). Scale Separation Reliability: What Does It Mean in the Context of Comparative Judgment? Applied Psychological Measurement, 42(6), 428–445. Retrieved from https://doi.org/10.1177/0146621617748321

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. Retrieved from https://doi.org/10.1080/0969594X.2019.1602027

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8:3, 370. Retrieved from https://www.rasch.org/rmt/rmt83b.htm

Appendix 1. Example rank ordering design



				Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7	Judge 8	Judge 9	Judge 10	Judge 11	Judge 12	Judge 13	Judge 14	Judge 15	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7	Judge 8	Judge 9	Judge 10	Judge 11	Judge 12	Judge 13	Judge 14	Judge 15	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7	Judge 8	Judge
ID	Year	Mark	Range	Pack 3	Pack 3	Pack 3	Pack 3	Pack 3	Pack 3	Pack 3	Pack 3	Pack 3	Pack 3	Pack 3	Pack 3	Pack 3	Pack 3	Pack 3	Pack 8	Pack 8	Pack 8	Pack 8	Pack 8	Pack 8	Pack 8	Pack 8	Pack 8	Pack 8	Pack 8	Pack 8	Pack 8	Pack 8	Pack 8	Pack 7	Pack 7	Pack 7	Pack 7	Pack 7	Pack 7	Pack 7	Pack 7	Pac
A15	18	4	1	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
B51	19	4	1	0	0	0	0	0	1	0	1	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
A60	18	5	1	0	0	1	1	1	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
B47	19	5	1	0	0	0	0	1	0	0	0	1	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
A10	18	6	1	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
B29	19	6	1	1	1	1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
A11	18	7	1	1	1	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
B09	19	7	1	0	0	1	0	1	0	1	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
A57	18	8	2	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
B04	19	8	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	
A52	18	9	2	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	
B10	19	9	2	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	1	1	0	0	0	0	0	
A58	18	10	2	0	1	0	1	1	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
B25	19	10	2	0	1	0	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
A54	18	11	2	0	0	1	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	
B44	19	11	2	1	0	0	1	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
A51	18	12	3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
B23	19	12	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	
A19	18	14	3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
B33	19	14	3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	
A16	18	15	3	0	0	0	0	0	0	0	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	
B32	19	15	3	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
A27	18	16	3	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	
B34	19	16	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	
A53	18	19	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
B46	19	20	4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
A36	18	22	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	
B20	19	23	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
A40	18	25	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	
A02	18	26	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	
B08	19	26	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	
B02	19	27	4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	
A28	18	28	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	
B36	19	28	5	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	
A46	18	29	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
B37	19	29	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	
A12	18	30	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	
B19	19	30	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
A04	18	31	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	
B53	19	31	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	
A05	18	32	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
B50	19	32	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
A06	18	33	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
B35	19	33	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	

Appendix 2. Task instructions

Rank ordering

Please read the whole instructions sheet and the FAQs carefully before starting the task

The purpose of this research

Ofqual has had a long term goal of making sure that awarding is valid and fit for purpose. As part of this, we want to see whether there are different methods of capturing expert judgement which could form part of the overall awarding process and maintaining standards from year to year. At the moment, expert judgement is directly influenced by statistics/outcomes in awarding – for example by the choice of scripts which are viewed by awarders. In addition, the scripts from the live session are normally considered without directly comparing them with those from the previous year, even in longer standing or 'settled' specifications, thus using a form of absolute judgement. Research around judgements indicates that humans are better (more accurate, more reliable) in making relative judgements (comparisons, 'x is better than y') than absolute judgements (this is/is not 'x'). For these reasons, it is possible that the current method of capturing expert judgement in awarding may not be the best method to detect genuine changes in student performance (improvements or deterioration) from year to year.

We are investigating different methods of capturing expert judgements which, unlike traditional awarding judgements, are (a) independent of statistics and (b) utilize relative (comparative) judgement. The method we are exploring in the current study involves rank ordering scripts – from the live session and the previous session, in terms of overall quality. When all the judgements from a number of judges are combined, and a scale of script quality derived from their judgements, it is possible to 'equate' any mark from one year to its equivalent mark in the following year using this script quality scale as the link.

This study builds on previous research using rank ordering and helps us further understand the functioning of this method, the extent to which the method yields consistent judgements, and how well it might work in a live operational context.

Materials

You have been sent in the post:

- Summer 2018 and summer 2019 question papers for WJEC GCSE English language specification (units C700/U1 and/or C700/U2)
- Two recording forms to capture your view of the relative difficulty of the 2018 vs. 2019 question papers (one form to compare the C700/U1 question papers from 2018 and 2019; one form to compare the C700/U2 question papers from 2018 and 2019)
- A set of 4 packs of scripts from 2018 and 2019 for Paper 1 (unit C700/U1), with the corresponding recording form in each.
- A set of 4 packs of scripts from 2018 and 2019 for Paper 2 (C700/U2), with the corresponding recording form in each.
- Two labelled envelopes to return all the materials to us in.

We have also emailed you:

- 2018 and 2019 mark schemes
- An excel file to record your rank ordering judgements on for Paper 1
- An excel file to record your rank ordering judgements on for Paper 2
- A link to the feedback questionnaire to complete after you have finished the rank ordering task

Instructions

1. Familiarising yourself with the assessment materials

Before attempting this exercise, please (re)-familiarise yourself with the question papers and specification from each administration. Please try to form a judgement about the relative difficulty of the two question papers. Please do this for each unit separately, i.e. compare question paper 1 from 2018 with question paper 1 from 2019, and question paper 2 from 2018 with question paper 2 from 2019. You may use your notes or the reports already produced.

You will need to take any differences in difficulty between the papers from different sessions into account when carrying out the rank ordering exercise.

Once you have formed an opinion regarding the relative difficulty of the papers, please record this on the relevant difficulty recording form.

2. The packs

You have two sets of packs of scripts:

- one set with 4 packs of summer 2018 and summer 2019 scripts for Paper 1
- one set with 4 packs of summer 2018 and summer 2019 scripts for Paper 2

There are six scripts in each pack, three from 2018 administration and three from 2019 administration. The scripts have been cleaned of student identifiers, marks and most marker annotations. You will notice that there are some missing pages. This is

because we removed any pages where candidates did not write anything, as well as any administrative pages to save on printing costs.

Some packs contain scripts which are moderately close to one another when marked conventionally, while others might contain scripts with a slightly greater range of quality. In general (though not always), the earlier packs contain higher quality scripts.

You should make no assumptions about the way in which the scripts are ordered within each pack. They are deliberately randomised.

The script labels do not relate to script total marks and were randomly generated.

3. The rank ordering task

Consider one pack at a time. For each pack, place the scripts into a single rank order, from best (rank 1) to worst (rank 6), based on a ***holistic judgement of overall quality***. Please take into account any differences in difficulty between the papers.

You do not have to complete all the packs in the same sitting – if you feel like it, you might want to break the task up into more manageable chunks.

The task should be carried out once for each pack of scripts. Do not to consider scripts from different packs at the same time, or compare scripts from paper 1 with those from paper 2 – scripts from different packs have to be kept separate.

4. Making the judgements

For each pack, you should endeavour to make a holistic judgement about each script's quality and its overall merit, relative to the other scripts in the pack, taking into account differences in difficulty between the two papers. You may use any method you wish to do this based on scanning the scripts and items and using your judgement to summarise the relative merits (see FAQs). You may wish to work in an environment where you have space to physically arrange the scripts into the rank order.

No tied ranks are allowed. If you are concerned that two or more scripts are genuinely of exactly the same quality you may indicate this by placing an equals sign (=) next to them on the recording form, but you must enter every script onto a separate line of the recording form.

Whilst it can be difficult to make relative judgements about scripts from different examinations, and with different knowledge and skills profiles, we ask that you do this as best you can, forming a holistic judgement of each script and using your own professional judgement to allow for differences in the exam papers.

You must take account of the whole work of each student. It is vitally important for the success of the research exercise that your judgment is based upon a holistic evaluation of each script. Please do not be tempted to base your judgments upon just one question or a subset of questions. Please consider all the responses that each student gave, and try to come to a view on the quality of the student's work relative to that shown in the other scripts in the pack.

Please do not collaborate with any of your colleagues who are completing this exercise as it is important that we have independent responses to the tasks. We are interested in your personal judgement about the quality of the scripts. Additionally, your colleagues will have a different combination of scripts in different packs.

If you have any uncertainties about what you are doing at any point in the process, please get in touch and we will be happy to talk you through it.

5. Use of mark schemes

We have emailed you the mark schemes for reference only – e.g. if you do not know a correct answer for a specific question. They are not to be used as the basis for the rank ordering.

You must not mark the scripts. You need to make an overall (or holistic) judgement about the quality of the scripts.

6. Recording your judgements

Once you have decided upon a single rank order for the scripts from a pack, please record the order on the recording form enclosed in the pack using the script ID's provided, and return the scripts to the pack before beginning another pack. The script ID is located at the top front of each script.

Please also answer the question about how easy or difficult it was to rank order questions in the pack by ticking the appropriate box.

Once you have completed all the packs, please fill in the appropriate electronic version of the recording sheet too. This is to ensure that we have access to the data as soon as possible.

7. Feedback questionnaire

After completing the rank ordering exercise, please fill in the feedback questionnaire using the link we emailed you.

8. Returning the materials

Please return the scripts and the recording sheets using the labelled envelope provided. Please put them in the post and email us the electronic version of the recording sheet **no later than Monday 15th July, 2019**.

Please also complete the **feedback questionnaire** via the link provided **by the same date**.

Please **email the electronic version of the recording form** to English.Language@ofqual.gov.uk

FAQs

What is rank ordering?

Rank ordering is a technique for capturing expert judgement for the purpose of comparing standards between different examinations (e.g. summer 2018 and summer 2019). Previous research exercises have found that rank ordering is a valid method for comparing standards between examinations. Essentially, a sample of scripts from two or more examinations are rank ordered by multiple judges (examiners, subject experts). These rankings are then analysed to place each script onto a single scale of quality. By looking at how the marks and grades from each examination are distributed on this scale we can map the performance standards between the examinations from different sessions and see if performance standards have changed or remained the same.

What should I do with the scripts in each pack?

Your main task is to rank order the scripts in each pack into a single rank order, from best (rank 1) to worst (rank 6) on the basis of script quality, allowing for any differences in difficulty between the papers from two examination sessions. Record your judgements on the recording sheet and return the scripts to the pack.

How should I arrive at a rank order?

You should make a holistic judgement of the quality of each script. As an experienced examiner, you probably have an understanding of what constitutes a good quality script for this specification.

We know it is a challenging task but it is really important that you do not refer to the mark scheme or mark the paper.

Different judges use different procedures and you may determine your own procedure. Some judges like to attach a very brief note, as a form of script summary or ‘aide memoir’ to some scripts (e.g. ‘good on X but less convincing on Y’) after reading/scanning to help them in the final consideration of script order.

Will marking the scripts help me?

No. In fact, it will work against the objectives of the exercise. Because mark scales for different specifications are not identical (e.g. a mark of 30 on one examination may not represent identical performance standards to a mark of 30 on the other

examination), marking the scripts will not help us place the two sets of scripts on a single scale. This can only be done by making holistic judgements about the quality of each script relative to the other scripts.

What if the tests from different examination sessions are of different difficulty?

Please try as best as you can to allow for this when making your rank ordering judgements.

Is there a 'right' answer to the order of the scripts?

This is not a 'test' whereby the researchers know the right answer and want to see if you can get it right! The 'right' order of scripts in any pack is the order that you determine by making a holistic judgement about the quality of each script relative to the other scripts in the pack.

Should I complete the whole task in one go?

You can work flexibly to fit around other commitments. There is no need to complete the whole task in one sitting.

How long should each pack take me?

Gradually as you become accustomed to this task you will no doubt speed up. We anticipate that each pack will take approximately 30 minutes in this context. Remember that the aim is to make holistic, intuitive judgements. Read each script, think about which are better or worse and put them in order. Try not to dwell on your decisions for too long.

What should I do if I have any questions?

Feel free to get in touch with us at any time!

Please contact Milja Curcin on English.Language@ofqual.gov.uk

PCJ

Please read the whole instructions sheet and the FAQs carefully before starting the task

The purpose of this research

Ofqual has had a long term goal of making sure that awarding is valid and fit for purpose. As part of this, we want to see whether there are different methods of capturing expert judgement which could form part of the overall awarding process and maintaining standards from year to year. At the moment, expert judgement is directly influenced by statistics/outcomes in awarding – for example by the choice of scripts which are viewed by awarders. In addition, the scripts from the live session are normally considered without directly comparing them with those from the previous year, even in longer standing or 'settled' specifications, thus using a form of absolute judgement. Research around judgements indicates that humans are better (more accurate, more reliable) in making relative judgements (comparisons, 'x is better than y') than absolute judgements (this is/is not 'x'). For these reasons, it is possible that the current method of capturing expert judgement in awarding may not be the best method to detect genuine changes in student performance (improvements or deterioration) from year to year.

We are investigating different methods of capturing expert judgements which, unlike traditional awarding judgements, are (a) independent of statistics and (b) utilize relative (comparative) judgement. The method we are exploring in the current study involves comparing pairs of scripts from two different examination sessions in terms of overall quality. When all the judgements from a number of judges are combined, and a scale of script quality derived from their judgements, it is possible to 'equate' any mark from one session to its equivalent mark in the following session using this script quality scale as the link.

This study builds on previous research using rank ordering and comparative judgement and helps us further understand the functioning of this method and the extent to which it yields consistent judgements, and how well it might work in a live operational context.

Materials

We have emailed you:

- Summer 2018 and summer 2019 question papers for the WJEC GCSE English language specification (components C700/U1 and C700/U2)

- Two recording forms to capture your view of the relative difficulty of the 2018 vs. 2019 question papers (one form to compare the C700/U1 question papers from 2018 and 2019; one form to compare the C700/U2 question papers from 2018 and 2019)
- A link to the online comparative judgement task for paper 1 (C700/U1)
- A link to the online comparative judgement task for paper 2 (C700/U2)
- 2018 and 2019 mark schemes

Instructions

1. Familiarising yourself with the assessment materials and deciding on paper difficulty

Before attempting this exercise, please (re)-familiarise yourself with the question papers and specification from each administration. Please try to form a judgement about the relative difficulty of the two question papers. Please do this for each component separately, i.e. compare question paper 1 from 2018 with question paper 1 from 2019, and question paper 2 from 2018 with question paper 2 from 2019. You may use your notes or the reports already produced.

You will need to take any differences in difficulty between the papers from different sessions into account when carrying out the comparative judgement task.

Once you have formed an opinion regarding the relative difficulty of the papers, please record this on the relevant difficulty recording form.

2. How do I access the judging software?

You will carry out the comparative judgment task using No More Marking judging software.

The links we sent to you takes you to a screen where you need to enter your email address and name. The study is set up with the email you used to communicate with us prior to this so please use the same email address to log in. You will then go straight into your allocation of judgements.

3. What are the technical requirements of the system?

No software installation is required. Being browser-based, the No More Marking judging system will run on any web browser on PC or Mac, although Internet Explorer seems more prone to slow loading than other browsers (but it should still work). Sometimes the files may not load fully, one (or both) may be blank or partially so – refresh your browser (hit F5) if this happens – refreshing has no ill effect on the judging software, it just reloads the same items, and you can do so as many times as you need. You will need to have a reasonably good internet connection, as each script file is over a thousand kb and two must be loaded for each judgement.

Ideally you would want to use a reasonably big widescreen monitor to see the scripts clearly side by side; however you can zoom in using normal controls or click on a script to see it larger in a separate tab, so it will still be practical to use a smaller screen.

4. The comparative judgement task

Once you access the task, you will be presented with pairs of scripts side by side on your computer screen (see image below) and should decide which is better based on a holistic judgement of overall quality.

The prompt at the top of the screen will say

'Which of these two scripts is better, based on a holistic judgement of overall quality?'

There will be 69 comparisons for you to make for each component.

The scripts in each pair could be from the same examination session, or one could be from 2018 and the other from 2019. The scripts have been cleaned of student identifiers, marks and marker annotations. The script labels do not relate to script total marks and were randomly generated.

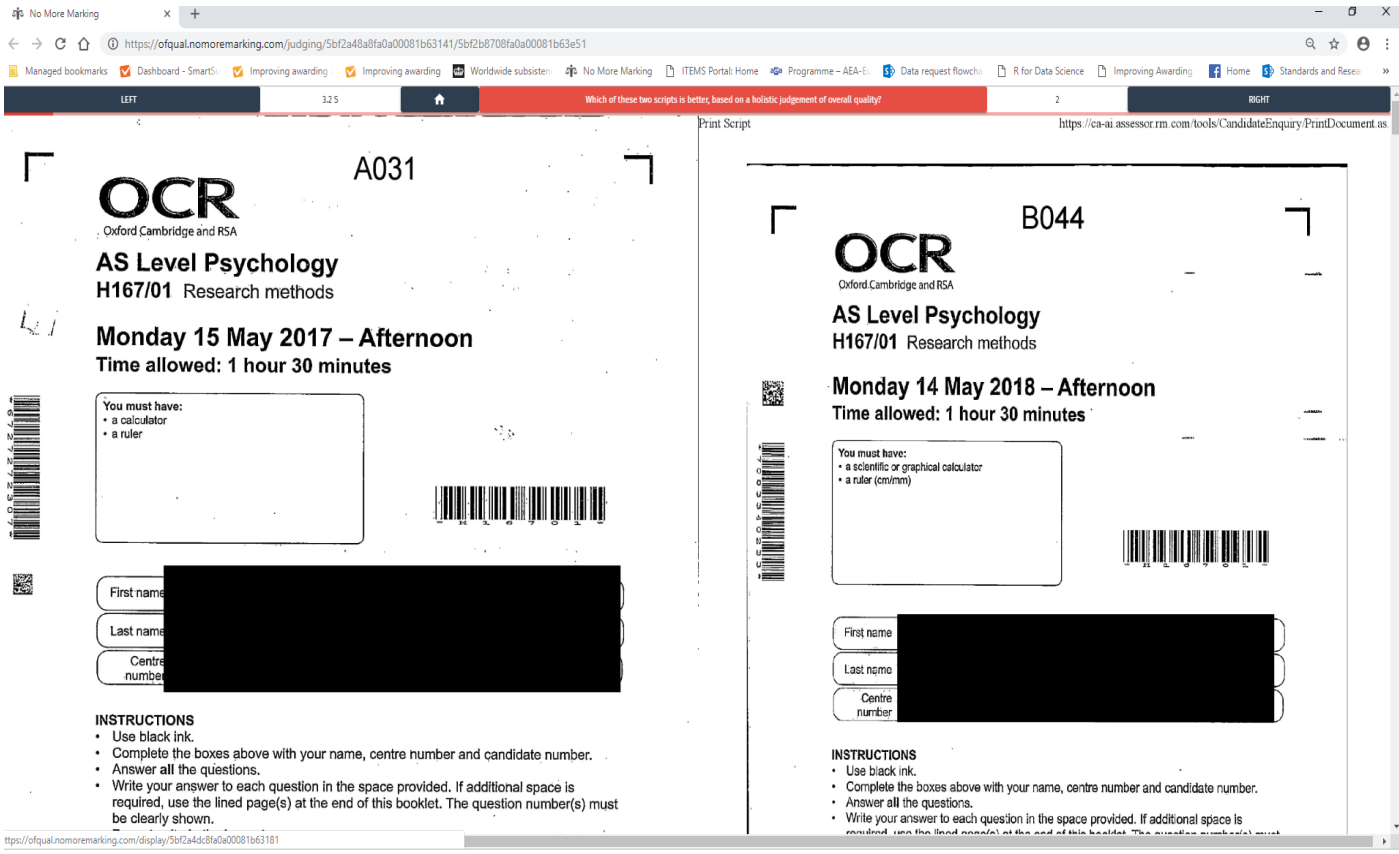
Consider each pair at a time. For each pair, make a decision on which of the two scripts is better, based on ***a holistic judgement of overall quality***. Please allow for differences in difficulty between the papers when making the quality judgements.

Once you have made a decision of which script is better, you should click on the 'Left' or the 'Right' button the top of the screen to indicate which script you think is better (see image below).

Ties are not allowed; you must pick one script even though you may feel that they are almost identical in quality.

You can look at both scripts side by side. Alternatively, you can click on each script and they will open in new individual tabs. Please make sure that you close any already opened individual tabs before you move on to the next pair of scripts.

You can close down the browser/tab at any time and pick up exactly where you left off by clicking on the link you will have been sent in an email by the No More Marking software. Therefore, you can work flexibly around your other commitments.



5. Making holistic judgements

For each pair, you should endeavour to make a holistic judgement about each script's quality and its overall merit, relative to the other script, taking into account differences in difficulty between the two papers. You may use any method you wish to do this based on scanning the scripts and items and using your judgement to summarise the relative merits. The aim is to make holistic, intuitive judgements - try not to dwell on your decisions for too long.

Whilst it can be difficult to make relative judgements about scripts from different examinations, and with different knowledge and skills profiles, we ask that you do this as best you can, forming a holistic judgement of each script and using your own professional judgement to allow for differences in the exam papers.

You must take account of the whole work of each student. It is vitally important for the success of the research exercise that your judgment is based upon a holistic evaluation of each script. Please do not be tempted to base your judgments upon just one question or a subset of questions. Please consider all the responses that each student gave, and try to come to a view on the quality of one script relative to that shown in the other script.

Please do not collaborate with any of your colleagues who are completing this exercise as it is important that we have independent responses to the tasks. We are

interested in your personal judgement about the quality of the scripts. Additionally, your colleagues will have a different combination of scripts.

If you have any uncertainties about what you are doing at any point in the process, please get in touch and we will be happy to talk you through it.

6. Use of mark schemes

We have emailed you the mark schemes for reference only – e.g. if you do not know a correct answer for a specific question. They are not to be used as the basis for comparative judgments.

You must not mark the scripts. You need to make an overall (or holistic) judgement about the relative quality of the scripts.

7. Feedback questionnaire

After completing the judging exercise, please fill in the feedback questionnaire using the link we emailed you.

8. Deadline

Please aim to complete the task **no later than Saturday, 13th July, 2019**. This task has to be completed before you can start on the rank ordering task.

Please return by email completed paper difficulty recording forms by the same date.

(FAQs similar to the rank ordering ones)

Appendix 3. Data cleaning

Media studies

After the initial Facets run of the model, two misfitting observations with standardised residuals over greater than 9 were removed, and the model rerun. All further analyses were based on the parameters from the second run.

Seven scripts won or lost all their comparisons (2 from 2017 and 5 from 2018) and were excluded from mark-measure correlation and regression analyses.

English literature 1

For paper 1, after the initial Facets run of the model, misfitting observations with standardised residuals greater than 4 were removed, and the model rerun. All further analyses were based on the parameters from the second run. No observations were removed for paper 2.

One 2017 in paper 1 and one 2017 script in paper 2 lost all their comparisons. These scripts did not contribute to measure estimation of other scripts, and were also excluded from mark-measure correlation and regression analyses.

English literature 2

RO

For paper 1, after the initial Facets run of the model, misfitting observations with standardised residuals greater than 6 were removed, and the model rerun. All further analyses were based on the parameters from the second run. No observations were removed for paper 2.

Two 2017 and two 2018 scripts in paper 1, and one 2017 script in paper 2 won or lost all their comparisons. These scripts did not contribute to measure estimation of other scripts, and were also excluded from mark-measure correlation and regression analyses.

Teacher PCJ

For paper 1, two judges with high outfit mean squares, which were also more than two standard deviations away from outfit mean were removed from the analysis (ouffitMS of 3.24 and 2.66). For paper 2, one judge with high outfit and infit mean squares that were also more than two standard deviations away from their respective means were also removed (infitMS of 1.87 and outfitMS of 2.19). All further analyses were based on the parameters with these judges excluded.

Two 2017 and four 2018 scripts in paper 1 won or lost all their comparisons and, alongside two outlier scripts, were excluded from mark-measure correlations and regression analyses. In paper 2, two 2017 scripts and six 2018 scripts won or lost all their comparisons, and alongside one outlier script, were also excluded from the abovementioned analyses.

Pinpointing PCJ

No judges or observations were removed from the analysis as overall the fit of the model was satisfactory for each grade boundary data set.

Psychology 1

RO

For paper 1, after the initial Facets run of the model, misfitting observations with standardised residuals greater than 6 were removed, and the model rerun. All further analyses were based on the parameters from the second run. No observations were removed for paper 2. Overall fit of the model was satisfactory.

Two 2017 and two 2018 scripts in paper 1, and two 2017 and four 2018 scripts in paper 2 won or lost all their comparisons. These scripts did not contribute to measure estimation of other scripts, and were also excluded from mark-measure correlation and regression analyses.

PCJ

Overall fit of the model was satisfactory and there was no need to remove any observations to improve the fit.

Three 2017 and ten 2018 scripts in paper 1 and four 2017 scripts and nine 2018 scripts in paper 2 won or lost all their comparisons and, alongside two outlier scripts, were excluded from mark-measure correlations and regression analyses.

Pinpointing PCJ

No judges or observations were removed from the analysis as overall the fit of the model was satisfactory for each grade boundary data set.

Psychology 2

RO

For paper 1, after the initial Facets run of the model, misfitting observations with standardised residuals greater than 6 were removed, and the model rerun. All further analyses were based on the parameters from the second run. No observations were removed for paper 2.

Three 2018 scripts in paper 1, and one 2018 script in paper 2 won or lost all their comparisons. These scripts did not contribute to measure estimation of other scripts, and were also excluded from mark-measure correlation and regression analyses. Two 2017 P1 scripts were removed as outliers.

PCJ

Overall fit of the model was satisfactory and there was no need to remove any observations to improve the fit.

Nine 2017 and five 2018 scripts in paper 1 and seven 2017 scripts and five 2018 scripts in paper 2 won or lost all their comparisons and, alongside two outlier scripts, were excluded from mark-measure correlations and regression analyses.

English language 1

Observations with standardised residuals above absolute 4 were removed from analyses.

Nine scripts with imputed measures from paper 1 and six from paper 2 were excluded from mark-measure correlations and regression analyses.

English language 2

Observations with standardised residuals above absolute 4 were removed from analyses.

Two scripts with imputed measures from each of paper 1 and paper 2 were excluded from mark-measure correlations and regression analyses.

English language 3

Observations with standardised residuals above absolute 4 were removed from analyses.

Six scripts with imputed measures from paper 1 and five from paper 2 were excluded from mark-measure correlations and regression analyses.

English language 4

RO

Observations with standardised residuals above absolute 4 were removed from analyses.

One outlier script from paper 1 and three scripts with imputed measures from paper 2 were excluded from mark-measure correlations and regression analyses.

PCJ

Observations with standardised residuals above absolute 4 were removed from analyses.

One outlier script from paper 1 and four scripts with imputed measures from paper 2 were excluded from mark-measure correlations and regression analyses.

Appendix 4. Judge fit statistics

Media studies RO

Judge	InfitMS	OutfitMS
1	1.09	1.17
2	0.66	0.48
3	1.05	1.13
4	0.77	0.53
5	1.35	1.47
6	0.96	0.75
Mean	0.98	0.92
SD	0.22	0.36

English literature 1 RO

	P1		P2	
Judge	InfitMS	OutfitMS	InfitMS	OutfitMS
1	1.05	0.92	0.9	0.80
2	0.86	0.72	1.15	0.89
3	0.97	0.83	1.07	0.97
4	1.19	1.14	0.81	0.78
5	0.99	0.88	0.94	0.66
6	0.96	0.74	1.13	1.23
Mean	1.00	0.87	1.00	0.89
SD	0.10	0.14	0.13	0.18

English literature 2 RO

	P1		P2	
Judge	InfitMS	OutfitMS	InfitMS	OutfitMS
1	0.82	0.61	0.93	0.6
2	0.99	1.04	1.35	1.38
3	1.18	1.23	0.99	0.93
4	1.16	0.96	0.8	0.4
5	1.03	0.86	0.78	0.51
6	0.81	0.67	1.17	0.9
Mean	0.99	0.96	1.00	0.79
SD	0.14	0.24	0.20	0.33

English literature 2 Teacher PCJ

	P1		P2	
Judge	InfitMS	OutfitMS	InfitMS	OutfitMS
1	Removed	Removed	1.22	1.45
2	0.83	0.40	0.56	0.33
3	0.45	0.30	1.04	0.68
4	1.05	0.40	0.51	0.38
5	1.63	1.38	0.48	0.26
6	0.51	0.35	1.12	0.97
7	0.90	0.54	1.35	1.44
8	0.79	0.46	0.66	0.32
9	0.23	0.11	0.45	0.36

10	0.50	0.30	0.78	0.67
11	0.62	0.33	0.73	0.49
12	0.41	0.23	1.01	0.77
13	0.90	0.84	0.81	0.64
14	0.42	0.27	0.53	0.32
15	0.54	0.38	1.20	0.84
16	0.86	0.49	0.63	0.47
17	0.76	0.48	1.04	1.20
18	1.22	0.63	1.07	1.30
19	0.96	0.54	Removed	Removed
20	0.45	0.12	0.58	0.34
21	0.59	0.34	1.24	1.01
22	1.29	0.99	0.32	0.22
23	1.77	1.27	1.16	0.75
24	0.74	0.40	Removed	Removed
25	0.50	0.38	0.66	0.42
26	0.48	0.29	1.04	0.89
27	0.39	0.19	1.19	1.36
28	1.16	0.77	0.92	0.74
29	0.44	0.26	0.61	0.44
30	Removed	Removed	0.63	0.42
31	1.35	0.59	0.67	0.43
32	0.83	0.57	1.51	1.16
33	1.55	1.03	0.96	0.89
34	0.68	0.33	1.24	0.99
35	1.24	0.98	0.56	0.51
36	0.73	0.43	0.34	0.21
37	0.48	0.17	0.60	0.45
38	0.94	0.72	1.06	0.87
39	0.56	0.28	1.04	0.79
40	0.90	0.65	0.22	0.14
41	0.79	0.57	1.18	1.53
Mean	0.81	0.51	0.84	0.70
SD	0.37	0.30	0.32	0.39

English literature 2 pinpointing PCJ

	P1_A		P1_E		P2_A		P2_E	
Judge	InfitMS	OutfitMS	InfitMS	OutfitMS	InfitMS	OutfitMS	InfitMS	OutfitMS
1	1.05	1.01	0.84	0.73	1.01	0.94	0.87	0.62
2	0.70	0.63	1.24	1.55	1.01	0.89	0.73	0.55
3	0.93	0.82	0.78	0.58	1.04	0.92	0.98	0.73
4	0.95	0.84	0.66	0.64	1.35	1.76	0.63	0.48
5	1.13	1.13	1.14	1.12	0.74	0.66	0.72	0.55
6	0.77	0.71	0.86	0.64	1.25	1.25	1.02	0.83
7	1.40	1.47	1.12	0.95	0.68	0.49	0.71	0.56
8	0.72	0.71	0.99	0.76	0.68	0.53	1.27	1.01
9	1.05	1.06	1.07	0.97	1.10	1.28	1.21	1.16
10	1.06	0.99	0.81	0.68	0.42	0.30	1.30	1.48
Mean	0.98	0.94	0.95	0.86	0.93	0.90	0.94	0.80
SD	0.21	0.25	0.19	0.30	0.29	0.44	0.25	0.33

Psychology 1 RO

	P1		P2	
Judge	InfitMS	OutfitMS	InfitMS	OutfitMS
1	0.81	0.59	1.01	0.94
2	1.26	1.47	1.03	0.98
3	0.91	0.73	1.09	1.06
4	1.00	0.85	0.80	0.68
5	0.91	0.76	0.92	0.86
6	1.04	1.05	1.10	1.01
Mean	0.99	0.91	0.99	0.92
SD	0.15	0.31	0.10	0.12

Psychology 1 PCJ

	P1		P2	
Judge	InfitMS	OutfitMS	InfitMS	OutfitMS
1	0.92	0.54	0.96	0.57
2	0.95	0.54	0.81	0.50
3	0.92	0.53	0.77	0.48
4	0.90	0.52	0.91	0.48
5	0.65	0.34	0.78	0.47
6	0.67	0.32	0.72	0.31
7	0.63	0.30	0.57	0.30
8	0.58	0.29	0.60	0.29
9	0.47	0.24	0.58	0.27
10	0.32	0.14	0.50	0.26
Mean	0.70	0.38	0.72	0.39
SD	0.22	0.14	0.15	0.12

Psychology 1 pinpointing PCJ

	P1_A		P1_E		P2_A		P2_E	
Judge	InfitMS	OutfitMS	InfitMS	OutfitMS	InfitMS	OutfitMS	InfitMS	OutfitMS
1	1.43	1.51	1.12	1.85	1.36	1.43	1.46	1.29
2	1.16	1.17	1.68	1.80	1.27	1.25	1.18	1.00
3	1.16	1.09	1.07	0.94	1.05	1.16	0.94	0.83
4	1.02	1.04	1.07	0.86	1.05	1.01	1.08	0.78
5	0.84	0.87	0.93	0.83	1.11	1.01	0.87	0.65
6	0.90	0.84	0.77	0.76	0.87	0.82	0.84	0.64
7	0.85	0.81	0.92	0.73	0.76	0.74	0.86	0.63
8	0.90	0.80	0.75	0.64	0.79	0.74	0.83	0.54
9	0.86	0.78	0.68	0.59	0.77	0.68	0.67	0.47
10	0.65	0.60	0.43	0.37	0.70	0.57	0.61	0.45
Mean	0.97	0.95	0.94	0.94	0.97	0.94	0.93	0.73
SD	0.22	0.26	0.33	0.49	0.23	0.28	0.25	0.26

Psychology 2 RO

	P1		P2	
Judge	InfitMS	OutfitMS	InfitMS	OutfitMS
1	0.93	0.74	0.97	0.82
2	0.95	0.83	1.02	0.89
3	1.02	0.78	1.01	0.93
4	1.01	0.71	0.97	0.95
5	0.89	0.64	0.87	0.66
6	1.18	1.12	1.10	1.14
Mean	1.00	0.80	0.99	0.90
SD	0.09	0.15	0.07	0.14

Psychology 2 PCJ

	P1		P2	
Judge	InfitMS	OutfitMS	InfitMS	OutfitMS
1	0.97	0.63	1.26	1.06
2	0.81	0.39	1.09	0.89
3	0.74	0.35	1.10	0.69
4	0.64	0.28	0.76	0.51
5	0.61	0.45	0.81	0.48
6	0.61	0.29	0.74	0.41
7	0.57	0.27	0.54	0.29
8	0.55	0.33	0.55	0.28
9	0.52	0.23	0.53	0.26
10	0.43	0.19	0.39	0.21
Mean	0.64	0.34	0.78	0.51
SD	0.16	0.13	0.29	0.29

English language 1 PCJ

	P1		P2	
Judge	Infit MS	Outfit MS	Infit MS	Outfit MS
1	0.52	0.16	0.75	0.32
2	0.39	0.11	0.76	0.34
3	0.77	0.28	0.69	0.28
4	0.70	0.25	0.68	0.31
5	0.51	0.18	0.39	0.14
6	0.74	0.26	0.94	0.59
7	0.81	0.27	0.41	0.14
8	0.47	0.13	0.65	0.21
9	1.06	0.61	0.60	0.25
10	0.65	0.22	0.81	0.26
11	0.44	0.15	0.84	0.41
12	0.69	0.26	0.70	0.26
13	0.53	0.17	0.42	0.14
14	0.70	0.19	0.67	0.19
15	0.55	0.20	0.68	0.24
Mean	0.64	0.23	0.67	0.27

SD	0.17	0.11	0.16	0.12

English language 2 RO

	P1		P2	
Judge	Infit MS	Outfit MS	Infit MS	Outfit MS
1	0.82	0.57	0.97	0.74
2	0.93	0.50	0.82	0.49
3	0.96	0.75	1.02	0.68
4	0.83	0.56	0.69	0.43
5	1.02	0.75	1.13	0.90
6	1.06	0.81	0.86	0.63
7	0.80	0.51	0.89	0.55
8	0.77	0.51	1.29	0.99
9	0.94	0.61	0.95	0.63
10	1.00	0.73	0.86	0.48
11	0.91	0.73	0.98	0.70
12	0.92	0.62	1.17	0.99
13	1.04	0.63	1.30	0.97
14	1.08	0.78	0.79	0.52
15	1.18	0.76	0.92	0.55
Mean	0.95	0.65	0.98	0.68
SD	0.12	0.11	0.18	0.19

English language 3 PCJ

	P1		P2	
Judge	Infit MS	Outfit MS	Infit MS	Outfit MS
1	0.42	0.14	0.79	0.36
2	0.69	0.31	1.29	0.48
3	0.40	0.10	0.43	0.15
4	0.90	0.26	0.29	0.08
5	0.78	0.31	1.11	0.60
6	0.38	0.11	0.69	0.28
7	0.41	0.16	0.78	0.37
8	1.58	0.84	0.39	0.12
9	0.37	0.12	0.47	0.17
10	0.75	0.29	0.86	0.29
11	0.60	0.34	0.59	0.26
12	1.56	0.76	0.72	0.26
13	0.43	0.13	0.65	0.22
14	1.22	0.41	0.80	0.32
15	0.55	0.18	0.81	0.33
16	0.50	0.21	1.17	0.51
17	0.66	0.23	0.96	0.36
18	0.60	0.21	0.74	0.44

19	1.64	0.71	0.86	0.33
20	0.74	0.26	1.01	0.66
Mean	0.76	0.31	0.77	0.33
SD	0.41	0.22	0.26	0.15

English language 4 RO and PCJ

	RO				PCJ			
	P1		P2		P1		P2	
Judge	Infit MS	Outfit MS	Infit MS	Outfit MS	Infit MS	Outfit MS	Infit MS	Outfit MS
1	0.53	0.26	0.84	0.45	0.71	0.31	1.12	0.61
2	0.83	0.42	0.85	0.59	1.04	0.46	0.87	0.58
3	0.92	0.46	0.84	0.46	0.53	0.18	0.90	0.33
4	0.61	0.31	0.61	0.29	0.70	0.23	0.76	0.21
5	0.65	0.41	0.79	0.49	0.92	0.48	0.62	0.28
6	0.80	0.78	0.83	0.45	0.78	0.30	0.51	0.16
7	1.12	0.82	0.93	0.67	0.72	0.43	0.60	0.26
8	0.70	0.38	0.78	0.38	0.81	0.28	0.70	0.27
9	1.02	0.48	0.80	0.42	0.78	0.28	0.66	0.24
10	1.09	0.50	1.60	1.08	0.83	0.41	0.71	0.35
11	1.13	0.56	1.16	0.86	1.15	0.74	0.87	0.57
12	0.94	0.50	0.80	0.56	0.82	0.34	0.82	0.37
13	0.66	0.43	0.80	0.42	0.61	0.20	0.82	0.37
14	1.00	0.46	0.83	0.43	0.81	0.42	0.81	0.34
15	0.89	0.55	0.65	0.31	0.65	0.31	0.68	0.22
Mean	0.86	0.49	0.87	0.52	0.79	0.36	0.76	0.34
SD	0.19	0.15	0.23	0.20	0.15	0.14	0.14	0.13

Appendix 5. Script statistics

In the tables below, the script IDs starting with 'A' denote scripts from Y1, and those starting with 'B' from Y2. The results in the tables are sorted by Measure, in descending order. The scripts where there are zeros in either Chosen or NotChosen columns had their measures imputed and were subsequently removed from regression analyses.

Media studies RO

Id	Measure	Measure SE	Infit	Outfit	Mark
B067	8.02	0.75	0.93	0.48	72
A061	6.21	0.5	1.11	0.87	74
B061	6.05	0.47	1.05	0.82	74
A030	5.73	0.5	1.03	1.18	67
A056	5.66	0.47	1.16	0.95	71
A062	5.37	0.6	0.82	0.56	63
B056	5.21	0.45	0.95	1.05	71
A011	4.98	0.44	0.97	0.75	69
A063	4.9	0.65	0.87	1.36	57
A067	4.79	0.43	0.76	0.69	72
A051	4.76	0.46	1.24	1.48	68
B062	4.7	0.47	0.82	0.55	63
A035	4.31	0.51	0.87	0.77	59
A007	4.26	0.46	1.21	1.24	61
B030	4.23	0.39	1.2	1.38	67
A031	4.11	0.44	1.23	1.62	66
B066	3.8	0.41	1.34	1.57	76
B051	3.78	0.4	1.1	1.24	68
A012	3.65	0.39	1.00	0.96	58
B002	3.61	0.4	0.74	0.66	77
B036	3.56	0.41	0.67	0.54	65
B035	3.48	0.45	0.76	0.58	59
A066	3.21	0.41	0.6	0.5	76
A036	3.18	0.42	1.08	0.97	65
A003	3.08	0.48	0.90	0.69	49
A069	3.06	0.47	0.70	0.45	53
A075	2.94	0.47	1.13	1.22	54
B001	2.84	0.45	1.14	1.32	62
A055	2.74	0.44	1.13	1.43	51
B012	2.7	0.43	0.61	0.45	58
A026	2.54	0.43	0.93	0.82	56
B021	2.52	0.45	0.79	0.66	52
A006	2.51	0.4	1.11	1.1	60
A021	2.42	0.43	0.86	0.77	52
B007	2.35	0.44	1.22	1.3	61
A005	2.32	0.43	0.87	1.01	64
A002	2.28	0.47	0.71	0.55	77
B070	2.24	0.48	0.98	0.69	43

B011	2.13	0.49	1.26	1.01	69
A050	2.07	0.46	0.91	2.01	44
B031	1.99	0.48	1.16	1.29	66
A039	1.89	0.5	0.95	0.6	48
A037	1.85	0.41	1.43	1.96	55
B006	1.8	0.44	1.3	1.29	60
A028	1.78	0.44	0.81	0.63	50
B069	1.4	0.41	0.69	0.61	53
B005	1.23	0.54	0.71	0.38	64
B063	1.17	0.45	1.56	2.65	57
B050	0.83	0.41	1.06	0.9	44
B026	0.78	0.48	0.9	0.68	56
B037	0.72	0.48	0.95	0.72	55
B028	0.58	0.46	0.92	0.77	50
B075	0.57	0.44	0.94	0.89	54
A009	0.52	0.41	1.03	0.9	46
A048	0.31	0.4	0.84	0.76	47
B020	0.3	0.43	0.91	0.72	45
B008	0.3	0.53	1.13	0.72	38
A020	0.29	0.44	1.09	0.99	45
B055	0.25	0.45	0.82	0.57	51
A029	0.23	0.45	1.01	0.88	41
B039	0.16	0.4	1.06	1.16	48
A081	0.09	0.46	0.88	0.58	42
A072	0.05	0.56	0.94	0.39	34
B048	-0.06	0.45	1.01	1.04	47
B003	-0.24	0.64	1.17	0.91	49
A004	-0.39	0.7	1.09	0.56	31
A013	-0.48	0.44	0.54	0.37	40
A070	-0.59	0.46	0.85	0.68	43
A047	-0.65	0.5	1.54	1.78	39
B009	-0.71	0.47	1.05	0.95	46
A027	-0.99	0.52	0.92	0.52	35
B071	-1.4	0.54	1.13	2.91	30
B081	-1.41	0.5	1.18	1.04	42
A068	-1.57	0.56	0.84	0.96	36
A073	-1.63	0.47	1.48	1.43	37
B029	-1.84	0.5	0.84	1.76	41
A014	-2.02	0.61	0.88	0.64	33
B073	-2.42	0.5	0.52	0.29	37
B013	-2.47	0.62	0.99	0.67	40
A032	-3.12	0.81	0.47	0.24	26
A080	-3.21	0.94	0.31	0.12	28
A071	-3.27	0.59	0.49	0.26	30
A008	-3.33	0.6	0.56	0.25	38
A065	-3.7	0.57	0.52	0.26	27
B027	-4.2	0.61	2.05	2.6	35
A045	-4.27	0.79	1.59	1.5	29

B047	-4.5	0.66	0.73	0.27	39
B057	-4.7	0.77	1.57	2.55	25
B068	-5.38	0.78	0.38	0.12	36
B072	-5.47	0.63	1.15	0.49	34
B014	-5.67	0.6	0.79	0.34	33
A044	-5.71	0.57	1.17	2.99	32
B044	-6.18	0.64	0.43	0.18	32
A023	-6.22	0.84	0.85	0.46	24
A042	-6.29	0.6	1.45	1.21	22
A022	-6.42	0.68	1.59	1.59	19
B042	-6.42	0.51	1.09	1.03	22
B065	-6.66	0.55	0.7	0.35	27
B045	-6.93	0.83	0.93	0.4	29
B023	-7.72	0.85	0.72	0.32	24
A059	-7.98	0.71	0.94	0.62	17
B032	-8.04	0.69	0.89	0.6	26
A016	-8.12	0.7	1.34	1.09	23
A057	-8.78	1.03	1.24	0.43	25
B016	-10.19	1.08	0.88	0.3	23

English literature 1 P1 RO

Id	Measure	Measure SE	Infit	Outfit	Mark
A065	6.77	1.15	1.48	0.59	70
B015	6.66	1.19	0.35	0.05	52
A052	4.01	0.62	0.84	0.51	53
B046	3.97	0.57	1.1	1.04	66
A029	3.85	0.59	0.95	0.61	54
B014	3.59	0.49	1.22	1.49	68
B009	3.58	0.51	1.21	0.96	70
A006	3.56	0.52	0.95	0.67	50
A046	2.87	0.43	1.06	0.92	68
B035	2.83	0.49	1.06	1.09	56
B029	2.76	0.49	1.07	0.89	51
B032	2.76	0.47	0.69	0.48	58
B008	2.73	0.45	0.91	0.88	55
B023	2.7	0.53	0.76	0.81	38
A062	2.62	0.47	0.92	0.7	52
B051	2.29	0.45	0.91	0.77	53
A005	2.15	0.47	0.87	0.73	66
A040	2.09	0.46	1.12	1.21	46
B017	2.08	0.43	1.01	0.84	62
A017	2.06	0.43	1.2	1.1	55
A066	1.96	0.46	1.31	1.13	60
B031	1.8	0.44	0.92	0.85	45
B012	1.72	0.44	1.28	1.57	49
A007	1.71	0.44	0.89	0.77	62

A016	1.67	0.54	0.84	0.63	37
B027	1.67	0.42	1.06	0.96	54
B004	1.58	0.45	1.46	1.68	64
A002	1.52	0.46	1.35	1.54	56
B021	1.51	0.43	0.94	0.8	57
B036	1.51	0.45	0.75	0.55	59
A015	1.21	0.46	0.75	0.68	43
A031	1.21	0.42	0.89	0.73	36
B024	1.2	0.55	0.79	0.42	32
B045	1.2	0.44	1.08	1.15	44
A042	1.14	0.51	1.07	0.81	35
A014	1.13	0.57	0.95	0.86	28
B039	1.08	0.47	1.05	0.92	37
A050	1.06	0.42	0.87	0.87	38
A067	1.05	0.41	0.82	0.73	40
A036	1.02	0.42	0.83	0.73	49
B041	1	0.48	1.48	1.21	50
A048	0.86	0.45	1	0.8	58
B016	0.85	0.42	0.87	0.76	39
A069	0.82	0.51	0.86	0.94	64
B037	0.71	0.55	1.13	0.66	30
A047	0.68	0.47	1.09	0.92	51
B025	0.64	0.43	1	0.94	40
A004	0.6	0.43	1.08	1.07	47
B044	0.59	0.42	1.27	1.2	42
B030	0.48	0.48	1.21	1.17	60
B048	0.45	0.46	0.98	1.03	36
A013	0.25	0.75	0.67	0.28	29
A060	0.17	0.43	1.15	1.25	44
A041	0.12	0.47	0.8	0.95	57
B006	0.09	0.59	0.7	0.45	25
B034	0.02	0.47	0.67	0.56	46
A038	-0.02	0.47	0.92	0.78	39
B001	-0.28	0.48	0.91	0.61	43
A024	-0.37	0.46	0.75	0.65	41
B028	-0.41	0.46	1.55	1.75	34
A010	-0.43	0.49	0.83	0.57	33
B022	-0.49	0.61	0.64	0.4	33
A012	-0.58	0.48	1.16	1.01	42
B003	-0.64	0.5	0.81	0.62	48
A056	-0.72	0.45	1.11	0.96	32
A020	-0.82	0.49	1.64	1.54	45
A064	-0.84	0.53	1.15	1.04	48
A051	-0.85	0.71	0.82	0.43	24
B043	-0.9	0.48	0.83	0.71	35
A033	-0.96	0.75	1.11	1.01	59
B011	-1.1	0.55	0.61	0.39	47
B019	-1.12	0.62	1.43	0.99	20

A001	-1.29	0.45	0.74	0.59	31
A043	-1.29	0.58	1.09	0.86	25
A021	-1.41	0.53	1.17	0.82	34
B002	-1.6	0.46	1.36	1.54	26
B033	-1.67	0.51	0.71	0.47	29
B050	-1.78	0.52	1.08	0.87	22
B020	-1.79	0.46	0.85	0.61	31
B013	-2.14	0.49	1.76	2.33	27
B038	-2.26	0.56	0.9	0.57	41
B049	-2.33	0.6	0.39	0.3	28
A022	-2.68	0.59	1.05	0.88	20
B005	-2.68	0.55	0.98	1.02	18
A053	-3.04	0.5	0.99	1.11	18
B047	-3.05	0.55	1.07	1.02	21
B007	-3.23	0.55	1.06	1	16
B010	-3.23	0.49	1.11	0.83	24
A034	-3.36	0.5	0.86	0.81	16
A037	-3.61	0.63	0.54	0.39	26
B042	-3.66	0.46	1.05	0.91	23
A068	-3.73	0.48	1.02	0.89	23
A009	-4.08	0.68	1.2	0.71	19
B040	-4.13	0.59	0.61	0.42	19
A044	-4.5	0.49	1.05	0.82	15
B018	-4.6	0.58	0.88	0.8	17
A070	-4.64	0.58	0.94	0.59	27
A058	-4.79	0.56	1.24	1.51	17
A071	-4.89	0.56	0.93	0.79	21
B026	-4.89	0.51	1.02	1.23	15
A027	-5.28	1.06	0.89	0.21	30

English literature 1 P2 RO

ID	Measure	Measure SE	Infit	Outfit	Mark
B046	4.21	0.66	1.03	0.91	44
B034	3.53	0.79	0.6	0.17	31
B033	3.52	0.59	0.87	0.7	41
B020	3.28	0.63	0.93	0.76	39
A039	3.18	0.59	0.92	0.71	34
B027	3.12	0.68	1.29	1.27	37
A032	3.05	0.61	0.93	0.7	33
A006	2.91	0.58	1.16	0.97	44
B048	2.81	0.58	0.8	0.46	28
A025	2.41	0.54	1.57	2.15	41
A004	2.07	0.48	1.26	1.35	35
A042	1.91	0.48	0.88	0.63	31
A026	1.76	0.56	0.67	0.54	42
B021	1.73	0.47	0.76	0.57	35

B038	1.57	0.82	0.99	1.54	18
A036	1.55	0.57	1.03	0.98	37
B043	1.54	0.54	1.58	1.23	22
B035	1.53	0.49	1.23	1.46	26
A041	1.48	0.56	1.29	1.24	22
B008	1.22	0.56	0.97	0.74	42
B017	1.21	0.6	0.68	0.44	23
A014	0.98	0.45	0.89	0.74	30
B031	0.96	0.55	0.73	0.58	33
B050	0.86	0.61	1.08	0.87	30
A019	0.79	0.59	0.97	0.7	39
A024	0.68	0.65	0.93	0.65	27
A020	0.65	0.49	1.08	1.02	32
A007	0.52	0.48	0.83	0.92	29
B019	0.4	0.53	1.5	1.43	32
B012	0.29	0.46	0.58	0.5	21
A002	0.28	0.54	0.63	0.49	18
A031	0.23	0.51	0.78	0.59	26
B044	-0.14	0.56	1.23	1.18	17
A015	-0.16	0.51	0.64	0.55	24
B036	-0.25	0.49	1.05	1.26	29
B028	-0.39	0.73	0.94	1.08	36
B007	-0.67	0.49	1.13	1.2	20
A016	-0.69	0.44	0.8	0.63	28
A023	-0.7	0.56	0.75	0.59	14
A029	-0.73	0.53	0.8	0.94	19
A034	-0.76	0.49	1.47	1.29	25
B002	-1.16	0.56	0.62	0.43	19
B040	-1.38	0.46	1.35	1.63	27
B011	-1.38	0.53	0.77	0.62	14
A012	-1.39	0.49	1.22	1.35	20
A028	-1.45	0.53	1.01	0.84	16
B001	-1.7	0.6	1.13	0.95	16
B006	-1.76	0.57	0.99	1.86	11
A022	-1.93	0.57	0.68	0.46	12
A021	-2.11	0.52	0.9	0.81	13
B041	-2.19	0.76	0.92	0.51	25
A001	-2.24	0.55	0.72	0.51	11
B016	-2.32	0.6	1.27	0.89	12
B049	-2.44	0.95	1.35	1.01	34
A033	-3.08	0.69	0.58	0.24	21
B037	-3.14	0.69	1.49	1.29	15
B022	-3.22	0.64	1.26	1.13	24
B039	-3.82	0.75	1.39	1.06	13
A040	-4.59	0.73	0.97	0.38	23
A043	-4.73	0.92	0.4	0.13	15
A018	-5.68	1.1	1.11	0.38	17

English literature 2 P1 – RO

Id	Measure	Measure SE	Infit	Outfit	Mark
B046	4.44	0.83	1.18	0.94	46
A008	4.38	0.82	0.87	0.6	48
A043	3.92	0.77	0.79	0.35	45
A015	3.21	1.08	1.22	1.03	23
A005	2.89	0.74	1.09	0.92	47
B035	2.45	0.82	0.86	0.34	31
A003	2.38	0.56	0.86	0.84	46
A022	2.38	0.72	0.79	0.46	29
A004	2.37	0.58	0.97	0.9	50
B004	2.28	0.58	1.41	1.9	48
B021	2.23	0.63	0.91	0.62	40
A033	2.06	0.75	1.32	0.83	38
A026	2.03	0.54	1.09	0.94	33
A028	1.84	0.58	0.72	0.55	34
B049	1.75	0.59	1.15	1.4	39
B020	1.73	0.59	0.55	0.39	43
A032	1.66	0.63	1.2	1.37	37
B034	1.28	0.56	1.13	1.13	36
A020	1.22	0.52	1.29	1.32	27
B005	1.17	0.62	0.58	0.44	49
B033	0.79	0.61	0.75	0.57	44
A034	0.74	0.54	1.02	0.81	39
B037	0.7	0.7	0.97	1.22	20
A042	0.69	0.6	1.15	0.97	44
B003	0.68	0.54	0.56	0.46	47
A010	0.67	0.56	1.1	1.16	49
A025	0.64	0.7	1.41	2.59	32
B050	0.64	0.56	1.17	1.18	35
A007	0.46	0.67	0.8	0.56	19
A029	0.36	0.55	0.85	0.7	35
B048	0.3	0.57	1.67	3.75	33
A036	0.2	0.65	0.76	0.54	40
B027	0.13	0.6	1.03	0.94	42
A021	0.07	0.61	1	0.96	28
A039	0	0.55	0.87	0.67	41
B019	-0.04	0.57	0.85	0.51	37
B008	-0.27	0.52	1.12	1.18	45
A014	-0.46	0.6	0.95	0.74	22
B001	-0.46	0.57	0.84	0.6	21
B028	-0.53	0.72	0.87	0.56	41
B011	-0.54	0.58	0.97	0.72	19
A040	-0.67	0.7	0.65	0.37	42
B016	-0.83	0.55	0.73	0.56	16

A019	-0.89	0.59	1.14	1.08	26
B036	-1.12	0.59	1.4	1.49	34
B044	-1.15	0.61	0.73	0.51	22
B043	-1.17	0.59	0.85	0.61	27
A009	-1.17	0.65	0.65	0.42	20
B038	-1.23	0.62	1.39	1.32	23
A001	-1.24	0.56	1.29	1.61	15
B017	-1.25	0.65	1.18	1.13	28
A018	-1.28	0.59	0.72	0.51	25
A016	-1.57	0.55	1.25	1.63	24
A031	-1.73	0.59	1.22	1.2	36
B031	-1.9	0.56	0.7	0.48	38
B039	-1.97	0.55	0.86	0.77	18
A006	-2.03	0.6	0.84	1.33	18
A002	-2.56	0.6	0.92	0.71	16
A023	-2.73	0.72	0.66	0.41	30
B002	-2.97	0.82	0.85	2.18	24
B012	-3.21	0.69	0.82	0.45	26
A024	-3.38	0.88	1.88	2.72	31
B040	-3.86	1.11	0.52	0.13	32
B006	-3.88	0.82	0.79	0.85	15
B022	-4.26	1.12	1.34	0.81	29
B007	-4.41	1.06	1.13	0.91	25

English literature 2 P2 – RO

Id	Measure	Measure SE	Infit	Outfit	Mark
A034	5.76	1.1	0.79	0.19	50
A015	5.46	0.79	0.93	0.47	47
A005	4.71	0.65	1.39	1.16	48
B016	4.54	1.04	0.98	0.54	44
B049	4.17	0.61	0.6	0.38	50
B048	3.77	0.62	1.35	1.25	49
A029	2.94	0.62	0.99	0.68	45
B044	2.74	0.72	0.75	0.4	40
B004	2.69	0.6	0.82	0.51	46
A026	2.4	0.67	0.92	1.58	49
B003	2.32	0.69	1.36	1.1	48
A008	2.26	0.56	0.96	1.01	46
A033	2.25	0.65	1.06	0.9	36
B012	2.14	0.79	1	0.66	34
B008	2.08	0.62	0.95	0.63	39
B039	1.96	0.66	1.03	0.8	32
A032	1.89	0.6	0.71	0.45	41
B038	1.47	0.53	1.14	1.13	47
A022	1.45	0.57	1.06	1.06	42
A020	1.44	0.61	1.05	0.72	26
A028	1.41	0.54	1.29	1.24	43

B020	1.2	0.6	1.17	1.41	30
B007	1.15	0.65	0.98	0.56	31
B041	1.04	0.59	0.9	0.6	29
A012	0.84	0.54	1	0.74	35
A036	0.76	0.79	0.91	0.6	24
A040	0.56	0.81	1.15	0.51	25
A014	0.5	0.58	1	0.88	40
A007	0.39	0.63	0.5	0.3	33
A019	0.32	0.6	1	0.69	34
B022	0.21	0.58	0.66	0.45	42
B002	0.11	0.66	2.55	6.01	27
A002	0.02	0.73	0.92	0.46	44
B021	-0.06	0.6	1.77	1.94	24
A031	-0.15	0.59	0.89	0.83	38
B050	-0.21	0.56	1.1	0.74	45
B019	-0.25	0.7	0.89	0.52	41
A043	-0.26	0.56	0.84	0.83	39
B043	-0.3	0.54	0.7	0.63	26
A042	-0.38	0.65	0.55	0.34	37
A001	-0.48	0.6	1.04	0.82	23
B034	-0.6	0.49	1.04	0.87	38
A018	-0.8	0.58	0.72	0.55	31
B036	-0.87	0.79	1.26	1.57	43
B031	-0.89	0.61	0.79	0.58	36
B033	-1.05	0.57	0.67	0.39	37
B027	-1.34	0.57	1.17	1.19	25
B040	-1.46	0.58	0.73	0.5	21
A009	-1.56	0.55	1.06	0.9	16
A039	-1.57	0.56	1.01	0.9	19
A024	-1.59	0.58	1.31	1.58	20
A021	-1.8	0.66	0.9	0.7	28
A004	-1.87	0.56	0.98	0.9	18
A010	-1.93	0.52	1.01	0.9	15
A023	-1.98	0.65	1.33	1.08	29
B017	-2.12	0.73	0.74	0.44	28
A041	-2.15	0.7	0.76	0.47	30
B009	-2.25	0.58	0.87	0.65	18
A016	-2.33	0.6	0.88	0.54	27
B006	-2.47	0.59	1	0.73	23
A025	-2.53	0.57	1.52	1.62	21
B046	-3.28	0.66	1.29	1.03	22
B001	-3.42	0.68	1.32	0.85	35
B035	-3.52	0.64	1.26	1	20
A003	-3.58	0.66	0.63	0.38	22
B037	-3.62	0.7	0.56	0.36	16
B028	-4.2	0.8	1.12	3.92	19
B011	-4.7	1.07	0.79	0.22	33
B005	-5.4	1.05	0.99	0.47	15

English literature 2 P1 – Teacher PCJ

ID	true.score	true.score.SE	infit	outfit	Mark	Chosen	NotChosen
B046	6.01	1.93	0.12	0.03	46	13	0
A032	4.07	1.36	0.29	0.08	37	9	2
A003	3.64	1.89	0.08	0.02	46	15	0
B003	3.46	1.16	0.29	0.10	47	11	1
A043	3.27	1.89	0.07	0.03	45	11	0
B033	3.18	1.94	0.13	0.05	44	7	0
A036	2.31	0.89	0.70	0.44	40	9	2
B020	2.11	1.05	1.27	1.07	43	9	1
A008	2.04	1.02	0.46	0.20	48	10	2
A041	1.79	0.99	0.32	0.17	43	9	2
A005	1.76	0.80	0.60	0.35	47	11	2
B021	1.70	0.83	1.12	0.74	40	9	3
B009	1.63	0.66	0.61	0.46	50	10	4
A025	1.62	0.91	0.94	0.69	32	12	2
A039	1.54	0.67	1.02	0.70	41	13	3
A034	1.50	0.88	0.70	0.41	39	7	3
A004	1.29	0.68	0.95	0.70	50	9	4
B034	1.25	0.67	1.17	1.06	36	9	4
B005	1.23	0.84	0.80	0.51	49	7	3
B048	1.09	0.66	0.98	0.82	33	9	4
A022	0.87	0.72	1.06	0.72	29	9	4
A028	0.82	0.61	0.84	0.71	34	8	6
B036	0.75	0.75	0.86	0.61	34	8	3
B004	0.56	0.87	0.53	0.31	48	7	4
B049	0.49	0.78	1.06	0.75	39	5	7
B037	0.46	0.65	1.09	0.97	20	7	6
B019	0.41	0.70	1.27	1.14	37	7	4
B039	0.39	0.68	0.95	1.10	18	6	7
B050	0.34	0.73	0.48	0.35	35	7	5
A040	0.26	0.77	0.85	0.52	42	6	7
A033	0.20	0.70	0.68	0.47	38	9	4
A026	0.18	0.65	1.21	1.10	33	7	6
A021	0.11	0.74	0.92	0.64	28	6	6
A012	0.03	0.75	0.89	0.65	21	6	5
B028	0.02	0.60	0.81	0.67	41	7	8
B008	-0.01	0.73	0.91	0.81	45	6	6
B027	-0.08	0.76	0.79	0.61	42	5	5
A029	-0.10	0.64	1.31	1.49	35	6	7
A020	-0.15	0.82	1.79	1.60	27	7	4
B038	-0.20	0.63	0.93	0.76	23	7	6
B035	-0.20	0.61	1.01	0.92	31	7	6
B031	-0.23	0.72	0.84	0.54	38	6	9
A010	-0.33	0.73	0.63	0.43	49	6	6
A001	-0.42	0.71	0.67	0.45	15	5	8

B001	-0.42	0.81	1.12	1.08	21	3	8
A015	-0.59	0.71	0.75	0.51	23	8	5
A042	-0.64	0.69	1.78	1.60	44	5	7
A019	-0.80	0.68	1.00	0.74	26	7	8
A007	-0.81	0.72	0.65	0.50	19	5	9
B007	-0.83	0.78	0.81	0.47	25	3	10
A014	-0.85	0.75	0.52	0.38	22	3	9
B017	-0.89	0.68	0.73	0.56	28	4	9
A031	-1.05	0.84	1.04	0.65	36	4	8
A009	-1.09	0.87	0.58	0.32	20	2	9
B022	-1.22	0.78	0.58	0.42	29	3	8
B012	-1.61	0.89	0.69	0.35	26	2	11
B043	-1.63	0.77	0.89	0.57	27	3	10
A018	-1.65	0.77	1.18	1.29	25	2	10
B041	-1.91	1.00	0.80	0.40	30	1	11
A023	-1.99	1.01	0.30	0.19	30	2	7
A016	-2.13	0.94	0.29	0.14	24	3	11
B002	-2.27	1.02	1.11	0.90	24	1	10
B006	-2.32	0.86	1.04	0.51	15	3	11
A006	-2.74	1.53	0.13	0.06	18	1	8
A002	-2.79	1.08	0.69	0.35	16	1	7
B040	-3.31	1.86	0.04	0.03	32	0	12
B044	-4.02	1.89	0.07	0.03	22	0	11
B016	-4.19	1.18	0.72	0.16	16	1	13
A024	-4.23	1.16	0.70	0.21	31	1	10
B011	-4.69	1.90	0.08	0.02	19	0	14

English literature 2 P2 – Teacher PCJ

Id	trueScore	true.score.SE	infit	outfit	mark	Chosen	Not.Chosen
A029	3.17	0.91	1.05	2.34	45	13	2
B016	2.53	0.75	0.73	0.47	44	9	3
B036	2.51	0.89	0.49	0.27	43	9	2
A002	2.47	0.80	1.40	1.48	44	8	3
A028	2.38	0.71	1.00	0.69	43	12	3
A005	2.19	0.75	0.71	0.40	48	11	3
A033	2.19	1.04	0.76	0.53	36	6	1
B003	1.97	0.81	0.62	0.43	48	7	3
B044	1.86	0.76	0.65	0.39	40	11	3
B033	1.86	0.85	0.49	0.27	37	10	2
A015	1.83	0.86	0.96	0.98	47	9	2
B049	1.67	0.93	0.97	0.78	50	9	2
B019	1.28	0.69	0.97	0.81	41	9	5
A012	1.24	0.87	1.03	0.53	35	9	3
A018	1.24	0.78	0.44	0.32	31	8	3
A034	1.16	0.72	0.87	0.77	50	7	4
A032	1.06	0.75	0.63	0.51	41	7	3
B041	1.01	0.80	0.74	0.61	29	7	4

B048	0.90	0.75	1.17	1.00	49	9	3
B038	0.84	0.63	1.24	1.27	47	8	6
A026	0.74	0.84	0.43	0.30	49	7	3
B002	0.63	0.71	1.12	1.18	27	9	3
B004	0.41	0.72	1.54	1.53	46	7	5
B012	0.35	0.74	0.62	0.53	34	5	5
A025	0.19	0.77	0.81	0.67	21	9	3
A031	0.16	0.82	1.24	1.34	38	7	2
A021	0.15	0.65	1.17	1.15	28	7	5
A014	0.07	0.70	0.67	0.58	40	6	5
A007	0.06	0.72	1.12	1.09	33	6	6
B017	0.01	0.75	0.67	0.48	28	3	8
B008	-0.04	0.66	0.48	0.38	39	7	7
A041	-0.14	0.72	0.63	0.42	30	7	6
B027	-0.19	0.91	1.91	1.88	25	4	4
B039	-0.20	0.61	0.74	0.57	32	7	8
A016	-0.22	0.76	1.34	1.37	27	3	7
B007	-0.27	0.62	0.89	1.07	31	6	8
B022	-0.31	0.70	0.71	0.61	42	4	6
A043	-0.31	0.65	1.38	1.25	39	6	7
B040	-0.33	0.64	1.45	3.60	21	6	7
A024	-0.34	0.65	0.96	0.84	20	6	7
A020	-0.43	0.80	0.87	0.75	26	2	7
A019	-0.44	0.70	0.88	0.75	34	4	7
B037	-0.44	0.71	0.66	0.53	16	5	6
A039	-0.54	0.68	1.25	1.19	19	4	8
A040	-0.55	0.69	0.60	0.43	25	5	8
B006	-0.60	0.79	0.76	0.54	23	4	6
A036	-0.61	0.86	0.70	0.37	24	3	8
B020	-0.66	0.60	1.02	1.04	30	6	6
A008	-0.66	0.70	1.00	0.94	46	5	5
B034	-0.71	0.61	1.23	1.24	38	6	6
B043	-0.71	0.63	0.87	0.69	26	6	8
B001	-0.93	0.72	0.72	0.58	35	4	7
A022	-1.00	0.63	0.77	0.56	42	7	8
A010	-1.16	0.81	0.57	0.36	15	4	7
B050	-1.18	0.71	0.72	0.61	45	4	7
A009	-1.19	0.65	0.88	0.74	16	4	8
A023	-1.32	0.83	1.08	1.06	29	3	7
B028	-1.34	0.81	1.28	1.19	19	2	8
B011	-1.45	0.71	0.63	0.50	33	3	8
A001	-1.57	0.78	0.96	0.80	23	2	9
A004	-1.73	0.73	0.89	0.69	18	3	8
B021	-1.79	0.80	0.93	0.80	24	2	11
B046	-1.82	0.86	1.54	1.32	22	2	8
A003	-1.91	1.00	0.92	0.63	22	1	10
A006	-2.11	0.89	0.73	0.53	32	2	12
A042	-2.28	1.03	0.55	0.26	37	1	9

B005	-3.58	1.88	0.06	0.03	15	0	10
B035	-3.81	1.88	0.05	0.03	20	0	11
B009	-4.03	1.90	0.08	0.04	18	0	8

English literature 2 pinpointing PCJ P1 grade A

ID	Measure	Measure.SE	Infit	Outfit	Mark	Chosen	NotChosen
A007	1.78	0.78	0.89	0.81	37	9	2
B049	1.56	0.85	1.16	2.26	39	8	2
B051	1.36	0.73	0.95	0.73	41	8	3
A010	1.34	0.71	1.04	0.95	37	8	3
A005	1.29	0.72	0.90	0.90	37	8	3
B067	1.10	0.82	0.83	0.72	37	8	2
B031	0.68	0.76	0.52	0.42	38	7	3
A002	0.43	0.66	1.16	1.18	37	5	5
B021	0.36	0.70	0.95	0.83	40	7	4
A032	0.26	0.64	0.88	0.85	37	6	5
A012	0.22	0.69	0.87	0.79	37	6	4
B064	0.22	0.69	0.84	0.80	39	6	4
B065	0.14	0.71	1.44	1.65	40	6	4
B019	0.10	0.68	0.98	0.86	37	6	5
B050	0.05	0.67	0.71	0.62	40	6	5
A004	0.04	0.61	0.99	0.97	37	5	6
B069	0.04	0.68	0.96	0.91	41	4	7
B057	-0.25	0.64	1.00	0.94	38	5	6
B028	-0.33	0.61	0.96	0.91	41	6	6
B060	-0.33	0.66	1.24	1.28	38	6	5
A003	-0.35	0.62	0.85	0.81	37	5	7
B063	-0.39	0.71	0.79	0.85	37	4	7
A014	-0.39	0.70	1.00	1.05	37	3	7
A013	-0.59	0.65	0.93	0.87	37	5	6
A001	-0.66	0.67	1.34	1.36	37	4	7
A008	-0.74	0.75	0.63	0.48	37	3	8
A006	-1.40	0.82	1.21	2.38	37	3	7
A009	-1.43	0.83	0.70	0.63	37	2	9
A011	-1.91	0.96	0.77	0.55	37	1	10
B053	-2.21	0.77	0.94	0.77	39	2	10

English literature 2 pinpointing PCJ P1 grade E

ID	Measure	Measure.SE	Infit	Outfit	Mark	Chosen	NotChosen
B052	3.15	1.03	0.56	0.24	18	10	1
A024	2.70	1.14	1.16	0.39	17	10	1
A021	2.05	1.05	0.93	0.41	17	10	1
A030	2.01	0.83	0.51	0.32	17	8	3
B068	1.84	0.78	1.22	1.09	20	8	3

A026	1.52	0.94	0.59	0.33	17	8	2
A027	1.05	0.82	0.54	0.35	17	9	2
B037	0.93	0.64	1.30	1.30	20	5	6
B056	0.55	0.70	0.65	0.62	19	6	5
A029	0.45	0.72	1.56	1.65	17	8	3
A034	0.13	0.74	1.00	0.76	17	5	6
A022	0.10	0.71	0.94	0.93	17	8	3
A036	0.04	0.77	1.40	1.45	17	6	5
A033	0.04	0.70	0.83	0.88	17	8	3
B070	-0.22	0.72	0.62	0.48	20	4	7
B001	-0.44	0.66	0.97	0.89	21	5	6
A031	-0.56	0.67	0.98	0.94	17	5	6
A028	-0.68	0.69	0.66	0.56	17	5	6
B058	-0.72	0.74	0.70	0.54	17	3	8
A023	-0.81	0.65	1.07	0.98	17	6	5
A035	-0.89	0.77	0.91	0.66	17	4	6
B054	-1.05	0.76	1.25	1.20	18	3	7
B039	-1.16	0.84	0.79	0.49	18	2	8
B061	-1.24	0.75	1.09	0.87	19	3	8
B066	-1.25	0.80	0.90	0.69	21	2	9
B055	-1.40	0.84	0.60	0.36	17	2	9
B059	-1.40	0.79	1.21	1.20	21	2	9
B062	-1.45	0.80	0.76	0.56	17	2	8
B011	-1.53	0.82	1.21	0.93	19	2	9
A025	-1.76	0.70	0.85	0.79	17	3	7

English literature 2 pinpointing PCJ P2 grade A

ID	Measure	Measure.SE	Infit	Outfit	Mark	Chosen	NotChosen
A013	2.47	1.00	0.53	0.26	39	10	1
B068	1.89	0.82	1.05	0.66	42	8	3
A007	1.18	0.65	1.34	1.36	39	8	4
A012	1.16	0.76	0.91	1.58	39	8	3
B052	1.14	0.77	0.77	0.59	43	8	3
B069	0.82	0.83	0.73	0.61	39	8	4
B072	0.82	0.71	1.04	1.01	39	7	3
A003	0.78	0.70	0.91	0.83	39	7	3
B008	0.44	0.81	0.91	0.68	39	5	5
B080	0.28	0.75	0.91	0.69	42	6	4
A014	0.12	0.65	1.05	0.99	39	6	5
A044	0.10	0.65	0.91	0.86	39	5	6
B019	0.01	0.71	1.50	1.30	41	5	7
B059	0.01	0.76	0.57	0.44	41	4	6
B044	-0.04	0.73	1.16	1.09	40	6	4
A002	-0.19	0.67	0.78	0.73	39	4	6
B050	-0.19	0.80	0.44	0.30	41	5	6

B036	-0.33	0.70	1.00	0.94	43	7	5
B070	-0.36	0.73	0.72	0.60	40	5	5
B051	-0.43	0.69	0.73	0.62	43	5	6
A010	-0.46	0.73	1.05	0.93	39	4	7
A005	-0.69	0.72	0.62	0.57	39	4	7
A008	-0.83	0.66	0.90	1.02	39	4	7
A043	-1.05	0.75	0.92	0.83	39	3	7
A006	-1.38	0.73	1.57	1.41	39	3	9
B065	-1.46	0.71	1.24	1.95	40	3	8
A009	-1.71	0.78	0.43	0.36	39	3	7
B022	-2.67	1.00	0.83	0.45	42	1	10

English literature 2 pinpointing PCJ P2 grade E

ID	Measure	Measure.SE	Infit	Outfit	Mark	Chosen	NotChosen
B058	2.69	1.06	0.61	0.23	19	10	1
A027	2.32	0.83	1.03	0.70	18	9	2
A022	1.98	0.84	0.70	0.52	18	9	2
B040	1.95	0.74	1.06	0.76	21	8	3
A025	1.70	0.79	0.66	0.44	18	8	3
B054	1.42	0.77	0.95	0.77	18	8	3
A021	1.41	0.71	1.22	1.06	18	7	4
A004	1.24	0.82	0.80	0.69	18	8	2
B063	1.05	0.71	1.04	1.21	21	6	5
A024	0.75	0.70	0.94	0.88	18	8	3
B067	0.73	0.70	1.08	0.98	22	5	6
B046	0.16	0.76	0.99	0.81	22	6	5
A033	0.08	0.72	1.46	1.65	18	7	3
A023	0.03	0.67	0.57	0.52	18	6	5
A029	0.01	0.69	1.03	0.96	18	5	6
B064	-0.18	0.72	1.03	0.89	19	5	6
A031	-0.22	0.68	0.74	0.63	18	5	6
B060	-0.27	0.67	1.07	1.00	22	7	4
B055	-0.33	0.83	0.75	0.44	18	4	7
B062	-0.39	0.79	1.02	0.78	21	4	6
B009	-0.45	0.69	0.99	0.91	18	4	6
A034	-0.70	0.75	0.78	0.55	18	3	8
A032	-0.81	0.73	1.41	1.50	18	3	7
B053	-0.84	0.82	0.53	0.34	20	3	8
B066	-0.98	0.81	0.54	0.34	20	7	4
A028	-1.20	0.71	0.86	0.73	18	3	8
B035	-1.33	0.98	0.84	0.55	20	2	8
B028	-1.72	0.96	0.90	0.69	19	2	9
A030	-4.02	1.86	0.04	0.03	18	0	11
A026	-4.06	1.87	0.05	0.03	18	0	11

Psychology 1 P1 – RO

ID	Measure	Measure SE	Infit	Outfit	Mark
A042	6.82	0.67	0.9	0.5	71
B065	6.81	0.53	1.3	1.11	67
A011	6.49	0.56	0.6	0.39	72
B035	6.27	0.49	0.88	1.16	71
A005	5.74	0.49	1.25	1.34	64
A056	4.9	0.44	0.6	0.44	73
A051	4.71	0.45	1.1	1.24	70
A030	4.6	0.43	1.2	1.32	68
B008	4.6	0.46	1.39	1.27	64
A031	4.52	0.52	0.7	0.44	66
B071	4.45	0.41	1.16	1.35	68
B004	4.08	0.51	0.72	0.47	59
B063	3.86	0.44	0.72	0.51	65
B012	3.76	0.46	1.26	1.62	62
A006	3.61	0.48	1.38	1.2	60
B042	3.58	0.46	1.01	0.87	57
A035	3.56	0.45	0.65	0.5	59
B057	3.52	0.52	0.92	1.14	56
A036	3.39	0.45	0.84	0.71	65
B021	3.12	0.44	1.25	1.42	45
A062	3.1	0.46	0.57	0.38	63
A063	3.06	0.48	1.33	1.85	57
B005	2.85	0.5	1.54	1.68	63
B006	2.84	0.44	0.87	0.79	60
B003	2.74	0.45	0.94	0.84	54
A026	2.73	0.42	0.82	0.67	56
B068	2.3	0.5	1.02	0.82	61
B080	2.29	0.49	1.49	1.43	47
A007	2.28	0.45	1.3	1.33	61
A037	2.24	0.45	0.84	0.7	55
B055	2.2	0.44	0.84	0.75	58
B026	2.14	0.39	0.98	0.92	55
B011	2.02	0.44	1.05	0.98	53
B013	2	0.45	0.68	0.5	49
A021	1.96	0.46	0.98	0.86	52
A001	1.84	0.43	0.7	0.6	62
A012	1.78	0.45	0.78	0.69	58
A069	1.78	0.45	0.89	0.71	53
A048	1.71	0.48	0.85	0.64	47
A020	1.52	0.44	1.17	1.07	45
B070	1.24	0.46	0.72	0.53	39
B056	1.16	0.48	1.05	0.93	44
B020	1.1	0.45	0.9	0.78	46
B023	1.1	0.47	1.17	1.09	52
A003	0.86	0.49	0.96	0.71	49

A055	0.86	0.45	1.14	1.37	51
B037	0.78	0.49	0.97	1.03	50
A009	0.75	0.45	0.83	0.7	46
B031	0.71	0.42	1.21	1.27	43
A075	0.69	0.5	1.19	1.41	54
A070	0.59	0.41	1.02	1.05	43
A028	0.48	0.45	1.03	1.48	50
A068	0.34	0.55	0.63	0.34	36
B045	0.18	0.44	1.48	1.48	38
A029	0.14	0.44	1.17	1.05	41
B044	0.01	0.52	1	0.64	51
B030	-0.18	0.45	1.37	1.87	34
B001	-0.21	0.46	0.9	0.76	41
B027	-0.33	0.46	0.68	0.5	48
A008	-0.36	0.45	1.01	0.88	38
B028	-0.43	0.43	1.03	0.88	40
A050	-0.54	0.46	0.59	0.44	44
A039	-0.63	0.44	0.81	0.72	48
A047	-0.92	0.45	0.91	0.81	39
A073	-0.97	0.42	0.76	0.61	37
B034	-1.11	0.44	0.98	0.77	37
A072	-1.22	0.46	0.92	0.69	34
B014	-1.23	0.52	1.32	1.13	42
B051	-1.28	0.46	1.14	1.27	28
B009	-1.39	0.44	0.79	0.64	30
A081	-1.46	0.49	1.17	0.97	42
A013	-1.48	0.48	1.34	1.43	40
A014	-1.51	0.47	1.21	1.07	33
B059	-1.56	0.48	1.08	0.8	29
A071	-1.58	0.54	0.95	0.71	30
B039	-1.86	0.44	0.84	0.72	33
A045	-1.93	0.49	1.03	0.76	29
A044	-2.28	0.43	0.91	0.75	32
A023	-2.31	0.65	1.09	0.58	24
B016	-2.62	0.58	0.65	0.34	36
B047	-2.67	0.45	0.53	0.42	32
B050	-2.71	0.48	1.15	0.82	23
B072	-2.73	0.49	1.5	2.02	26
A027	-2.76	0.45	0.84	0.7	35
B069	-2.81	0.48	1	0.76	31
A016	-2.96	0.5	1.11	0.7	23
A065	-2.99	0.51	1.38	0.97	27
B075	-3.21	0.46	1.01	1.28	25
B073	-3.28	0.44	0.86	0.75	24
A022	-3.29	0.53	1.21	1.24	19
A080	-3.47	0.47	0.7	0.49	28
B036	-3.78	0.54	1.27	1.25	35
A004	-3.83	0.48	0.9	0.62	31

B048	-4.01	0.54	0.69	0.38	27
A057	-4.24	0.47	0.87	0.72	25
B029	-4.51	0.51	1.83	2.34	19
B067	-4.82	0.55	0.73	0.57	21
B081	-5.34	0.54	0.84	0.91	20
A059	-5.48	0.54	0.88	0.87	17
B022	-6	0.54	1.34	1.9	16
B032	-6.08	0.57	0.65	0.38	22
A032	-6.58	0.66	0.43	0.17	26
B062	-6.95	0.54	0.79	0.71	15
A034	-7.55	0.65	1.08	1.02	14
A067	-8.22	0.75	1.43	1.26	16
B007	-9.09	1.11	0.76	0.11	17

Psychology 1 P2 – RO

ID	Measure	Measure SE	Infit	Outfit	Mark
A067	5.09	0.77	0.87	0.51	58
B048	4.62	0.4	1.04	0.98	65
B013	4.39	0.46	1.01	0.89	61
B027	4.31	0.64	0.95	0.63	52
B062	4.26	0.38	1.17	1.23	64
A055	4.25	0.39	0.97	1.06	67
A081	4.14	0.38	0.94	0.99	70
A009	3.83	0.43	0.8	0.72	61
A016	3.71	0.38	0.86	0.78	66
A056	3.6	0.43	0.97	0.86	63
B035	3.51	0.41	1.17	1.21	60
A005	3.35	0.46	1.08	0.83	55
A030	3.35	0.5	1.2	1.07	57
A020	3.28	0.43	1	0.98	65
B003	2.95	0.4	1.02	1.04	58
B055	2.93	0.41	0.81	0.77	57
A039	2.82	0.41	0.67	0.55	59
A007	2.79	0.44	0.7	0.56	56
B036	2.71	0.4	1.22	1.26	62
B005	2.65	0.42	1.17	1.27	59
A028	2.5	0.44	0.73	0.58	60
B044	2.48	0.42	1	0.91	56
A080	2.44	0.42	1.22	1.34	62
B023	2.3	1.08	1.12	0.37	31
B022	2.18	0.4	1.19	1.27	55
A031	2.08	0.45	1.12	1.44	53
B047	1.98	0.52	0.71	0.51	45
A071	1.91	0.43	1.34	1.72	54
B016	1.85	0.46	1.2	1.38	51
A048	1.83	0.44	0.89	0.88	64

A037	1.82	0.44	1.41	1.69	52
B001	1.82	0.44	0.84	0.74	53
B026	1.81	0.42	0.95	0.81	54
B056	1.75	0.43	0.79	0.63	47
B072	1.59	0.46	1.18	1.31	48
A068	1.52	0.85	0.75	0.34	36
A065	1.46	0.48	0.89	0.74	51
B081	1.24	0.48	0.76	0.7	44
A011	1.09	0.44	0.86	0.75	48
B012	1.06	0.48	1.06	1.41	46
A057	1.02	0.43	0.89	1.15	49
A062	0.89	0.47	0.81	0.94	37
A051	0.59	0.42	1.01	1.17	42
B037	0.46	0.43	0.72	0.64	39
B071	0.36	0.42	0.99	1.03	43
B009	0.29	0.44	1.2	1.24	36
B021	0.28	0.45	1.14	1.04	42
A070	0.13	0.43	0.86	0.75	45
B045	0.08	0.43	0.7	0.6	32
A008	0.04	0.5	0.92	0.77	50
A004	0.01	0.45	0.78	0.74	44
A012	-0.01	0.46	1.19	1.19	43
B028	-0.15	0.56	0.9	0.65	50
A069	-0.38	0.44	0.85	0.75	40
B073	-0.47	0.58	1.31	0.94	49
B034	-0.5	0.63	0.75	0.47	20
B006	-0.52	0.48	0.86	0.6	29
A036	-0.62	0.45	0.96	1.62	34
A042	-0.64	0.59	1.25	0.88	33
A026	-0.71	0.49	1.28	1.04	31
B031	-0.72	0.46	1.32	2.15	40
A050	-0.8	0.57	0.64	0.31	46
B068	-0.82	0.63	1.43	1.25	41
B057	-0.91	0.47	0.86	0.77	35
A034	-0.96	0.44	1.15	1.03	41
B004	-0.96	0.49	0.7	0.5	38
A021	-1	0.49	0.65	0.57	38
A032	-1.45	0.47	0.72	0.51	32
A014	-1.6	0.54	0.84	0.55	25
A063	-1.79	0.48	1.08	1.95	39
A013	-1.82	0.56	1.05	0.87	27
A073	-1.82	0.5	0.62	0.41	35
B008	-1.85	0.52	1.29	1.16	33
A059	-1.92	0.58	1.04	1.05	17
B067	-1.99	0.53	1.14	1.18	37
B042	-2.04	0.45	1.27	1.19	28
A035	-2.28	0.48	0.61	0.51	29
B032	-2.44	0.49	0.82	0.61	26

A003	-2.58	0.48	0.93	0.83	23
A006	-2.59	0.49	1.38	1.4	22
A001	-2.59	0.52	1.3	1.36	24
A027	-2.6	0.48	1.2	1.23	18
B039	-2.78	0.58	0.75	0.61	34
A045	-3.34	0.48	1.02	0.79	26
B014	-3.35	0.49	0.88	0.69	24
A023	-3.44	0.48	1.07	1.12	21
A047	-3.49	0.46	1.1	1.07	16
B059	-3.59	0.5	0.96	0.64	27
A022	-3.8	0.53	1.23	1.4	19
B075	-3.81	0.44	1.33	1.4	15
B051	-3.86	0.63	1.42	1.45	30
B065	-3.95	0.57	0.66	0.41	22
A029	-3.99	0.55	0.89	0.54	28
B070	-4.02	0.48	0.8	0.64	19
B069	-4.24	0.47	0.76	0.59	12
B063	-4.27	0.59	1.09	0.87	25
B050	-4.58	0.55	1.11	0.91	21
A075	-4.76	0.55	0.94	0.7	30
B030	-5.13	0.66	0.99	0.47	23
B020	-5.45	0.77	0.88	0.53	18

Psychology 1 P1 – PCJ

ID	Measure	Measure.SE	Infit	Outfit	Mark	Chosen	NotChosen
A011	5.26	2.01	0.21	0.04	72	10	0
B004	5.22	1.96	0.15	0.03	59	10	0
B008	5.21	1.94	0.13	0.03	64	10	0
B071	4.77	1.91	0.09	0.03	68	10	0
A042	4.66	1.88	0.07	0.03	71	10	0
B035	4.62	1.98	0.18	0.04	71	10	0
B012	4.03	1.33	0.23	0.09	62	9	1
A035	3.65	1.17	0.19	0.10	59	8	2
A030	3.39	1.00	0.96	0.76	68	9	1
A056	3.27	1.34	0.17	0.08	73	9	1
B065	3.00	1.05	0.86	0.40	67	9	1
A051	2.98	1.32	0.15	0.08	70	8	2
A031	2.98	0.91	0.56	0.32	66	8	2
A026	2.88	1.06	0.81	0.35	56	9	1
A005	2.61	1.02	0.86	0.47	64	9	1
B042	2.57	0.76	0.89	0.68	57	7	3
A003	2.52	1.03	1.05	0.62	49	8	2
A062	2.47	0.91	0.74	0.40	63	8	2
B003	2.33	0.84	0.44	0.30	54	7	3
A021	2.27	0.81	1.05	1.02	52	8	2
B006	2.23	0.91	0.36	0.23	60	6	4
B044	2.09	0.96	1.43	0.68	51	7	3

B009	2.08	1.02	0.32	0.17	30	7	3
A012	2.03	0.69	0.83	0.78	58	4	6
A006	1.97	1.27	1.22	0.33	60	9	1
B020	1.75	0.86	1.03	0.64	46	8	2
B021	1.75	1.02	0.33	0.17	45	8	2
B037	1.68	1.03	0.99	0.43	50	7	3
B055	1.66	0.95	0.38	0.22	58	7	3
B011	1.60	0.98	0.32	0.19	53	7	3
B063	1.55	0.93	1.51	1.13	65	7	3
A036	1.49	0.90	1.38	1.07	65	6	4
A050	1.43	0.89	0.72	0.42	44	5	5
B034	1.39	0.79	0.75	0.64	37	5	5
B005	1.35	0.90	0.32	0.22	63	7	3
B026	1.21	0.90	0.45	0.27	55	6	4
A063	1.19	1.00	0.55	0.26	57	7	3
B057	1.03	0.83	0.62	0.41	56	6	4
A001	1.02	0.80	0.77	0.51	62	6	4
B013	1.00	0.97	1.12	0.70	49	6	4
B031	0.99	1.04	0.92	0.45	43	6	4
B023	0.98	0.89	0.75	0.41	52	7	3
A037	0.88	0.86	1.09	0.69	55	5	5
A070	0.84	0.87	0.87	0.60	43	7	3
A020	0.77	0.95	0.63	0.32	45	6	4
A069	0.75	0.98	1.01	0.62	53	7	3
A007	0.66	0.93	0.63	0.33	61	5	5
A008	0.46	0.95	1.14	0.66	38	7	3
B045	0.46	0.88	1.53	1.16	38	6	4
A075	0.45	1.21	0.22	0.10	54	4	6
B080	0.39	1.08	0.34	0.16	47	8	2
A028	0.39	0.89	0.51	0.30	50	4	6
B059	0.36	0.74	0.59	0.48	29	4	6
B014	0.07	0.89	0.26	0.20	42	5	5
B030	0.04	0.93	0.44	0.25	34	4	6
A047	-0.08	0.82	1.14	0.80	39	5	5
A009	-0.11	0.92	0.57	0.34	46	5	5
A048	-0.15	0.97	0.37	0.21	47	5	5
A068	-0.16	0.89	0.55	0.32	36	4	6
B068	-0.17	0.92	0.56	0.31	61	4	6
A039	-0.35	1.03	0.23	0.14	48	5	5
B001	-0.36	0.97	1.30	0.82	41	5	5
B027	-0.41	0.79	1.13	0.82	48	4	6
B028	-0.42	1.15	1.84	1.02	40	3	7
A023	-0.58	1.03	0.22	0.14	24	6	4
B056	-0.70	0.77	1.00	0.93	44	5	5
A016	-0.92	0.90	1.70	1.32	23	5	5
A055	-0.96	1.05	0.18	0.13	51	6	4
B069	-1.15	0.82	0.40	0.29	31	4	6
B070	-1.17	0.89	0.44	0.28	39	5	5

A014	-1.24	1.05	0.39	0.19	33	4	6
A071	-1.32	1.37	0.18	0.07	30	2	8
A081	-1.47	0.86	0.75	0.51	42	4	6
A029	-1.59	0.96	0.93	0.48	41	2	8
A013	-1.69	0.95	1.24	0.76	40	5	5
A027	-1.71	1.03	0.22	0.14	35	4	6
B039	-1.73	0.91	1.17	0.70	33	2	8
A065	-1.73	1.31	0.17	0.08	27	3	7
A022	-1.95	1.13	0.53	0.20	19	3	7
A073	-1.96	1.15	0.26	0.12	37	2	8
A032	-1.99	1.31	0.20	0.08	26	1	9
B047	-2.03	0.79	1.01	0.79	32	4	6
A080	-2.16	0.88	0.72	0.41	28	4	6
B075	-2.32	1.06	0.48	0.21	25	1	9
B029	-2.47	1.13	0.63	0.23	19	2	8
A004	-2.52	1.02	0.42	0.20	31	2	8
B050	-2.60	0.88	0.94	0.56	23	2	8
B072	-2.67	0.89	0.45	0.28	26	2	8
B073	-2.67	0.93	0.37	0.22	24	2	8
A059	-2.74	0.93	0.98	0.54	17	3	7
A072	-2.79	1.02	0.62	0.28	34	2	8
A045	-2.83	0.99	0.74	0.35	29	2	8
B081	-2.85	0.88	1.36	1.12	20	3	7
B051	-3.20	1.04	0.52	0.24	28	1	9
A057	-3.27	1.06	1.02	0.49	25	1	9
A034	-3.34	1.07	0.68	0.28	14	1	9
B016	-3.52	1.93	0.12	0.03	36	0	10
A044	-3.54	1.47	0.21	0.07	32	1	9
B067	-3.95	1.59	0.20	0.06	21	1	9
B007	-3.99	2.00	0.20	0.04	17	0	10
B022	-4.04	1.53	0.16	0.06	16	1	9
B062	-4.41	1.88	0.06	0.03	15	0	10
B036	-4.74	1.89	0.08	0.03	35	0	10
B048	-4.80	1.89	0.08	0.03	27	0	10
B032	-5.20	1.92	0.11	0.03	22	0	10
A067	-5.96	1.93	0.12	0.03	16	0	10

Psychology 1 P2 – PCJ

ID	Measure	Measure.SE	Infit	Outfit	Mark	Chosen	NotChosen
B062	5.77	1.91	0.09	0.03	64	10	0
A016	4.90	1.88	0.06	0.03	66	10	0
B001	4.38	1.94	0.13	0.03	53	10	0
A081	4.05	1.94	0.13	0.03	70	10	0
A020	3.72	2.03	0.24	0.04	65	10	0
A080	3.49	1.41	0.10	0.06	62	8	2
B005	3.48	1.07	0.96	0.42	59	9	1
A005	3.38	1.66	0.12	0.05	55	9	1

A071	3.22	1.11	0.37	0.16	54	9	1
B048	3.00	1.00	0.99	0.81	65	9	1
A030	2.81	1.00	0.81	0.36	57	8	2
A057	2.66	1.04	1.17	0.86	49	9	1
B036	2.55	0.97	0.43	0.22	62	8	2
A068	2.54	1.28	0.30	0.11	36	9	1
B013	2.52	1.03	0.57	0.27	61	9	1
A009	2.47	1.31	1.72	0.50	61	9	1
B056	2.47	0.85	0.75	0.49	47	8	2
B027	2.37	1.19	0.15	0.09	52	6	4
A048	2.36	1.05	0.87	0.40	64	9	1
B044	2.26	0.91	1.29	1.10	56	8	2
A050	2.12	0.93	0.38	0.22	46	8	2
A056	2.03	1.17	1.21	0.41	63	9	1
A051	2.02	1.10	0.66	0.25	42	9	1
B026	1.76	0.79	0.49	0.38	54	5	5
B057	1.74	0.90	0.92	0.52	35	7	3
A004	1.72	0.94	0.43	0.24	44	7	3
A007	1.72	0.90	0.54	0.31	56	8	2
A055	1.71	1.04	0.52	0.24	67	9	1
B012	1.65	0.97	0.43	0.22	46	8	2
A067	1.64	1.07	0.76	0.33	58	7	3
B055	1.62	0.76	0.80	0.74	57	6	4
B047	1.59	1.08	1.30	0.79	45	9	1
A031	1.59	1.08	0.43	0.18	53	7	3
A028	1.50	1.21	0.23	0.10	60	8	2
B035	1.38	1.04	1.03	0.50	60	7	3
B072	1.38	0.95	0.88	0.48	48	7	3
A008	1.30	0.77	0.56	0.42	50	7	3
B003	1.27	0.91	0.78	0.44	58	7	3
A065	1.23	0.83	1.23	1.12	51	6	4
A039	1.10	0.93	1.25	1.06	59	6	4
B016	0.92	0.84	0.51	0.33	51	7	3
A037	0.83	0.76	0.81	0.66	52	7	3
B037	0.71	0.91	1.15	1.00	39	5	5
B022	0.65	0.93	0.34	0.21	55	5	5
A070	0.64	0.98	0.47	0.23	45	7	3
A014	0.31	0.86	0.89	0.66	25	5	5
B023	0.31	0.96	0.54	0.30	31	3	7
A011	0.30	1.14	0.69	0.25	48	5	5
A062	0.28	0.84	1.27	1.06	37	5	5
A042	0.27	0.81	0.96	1.09	33	5	5
A045	0.25	0.95	1.21	1.01	26	6	4
A063	0.23	0.83	0.73	0.51	39	6	4
B045	0.12	0.70	0.82	0.78	32	5	5
B081	0.02	0.99	0.45	0.23	44	4	6
B028	-0.02	0.81	0.68	0.47	50	3	7
B067	-0.02	0.89	0.55	0.33	37	2	8

B009	-0.06	0.83	0.91	0.65	36	6	4
B031	-0.11	1.00	0.22	0.15	40	4	6
A073	-0.18	0.89	0.51	0.32	35	5	5
A069	-0.21	0.92	1.36	1.01	40	5	5
A034	-0.35	1.00	0.30	0.17	41	5	5
A001	-0.35	0.87	0.41	0.27	24	6	4
A036	-0.57	0.99	0.32	0.18	34	5	5
B004	-0.59	0.77	0.58	0.45	38	5	5
B068	-0.65	0.89	1.10	0.98	41	4	6
B006	-0.67	0.91	1.43	1.03	29	3	7
A059	-0.79	0.93	0.63	0.36	17	3	7
B021	-0.91	0.96	0.24	0.17	42	4	6
B008	-1.19	1.13	0.21	0.12	33	2	8
A029	-1.24	0.90	0.90	0.51	28	4	6
B032	-1.30	0.79	1.07	0.98	26	4	6
B051	-1.35	0.92	0.54	0.30	30	4	6
B073	-1.45	1.10	0.60	0.23	49	4	6
A006	-1.66	1.13	0.86	0.30	22	3	7
A012	-1.68	0.96	0.90	0.44	43	2	8
B042	-1.76	1.03	0.45	0.21	28	3	7
B075	-1.79	0.95	1.45	1.14	15	2	8
A021	-1.80	0.85	0.67	0.41	38	4	6
B014	-1.87	0.94	1.28	0.70	24	2	8
B039	-1.94	0.88	0.80	0.44	34	3	7
B034	-1.97	1.22	0.98	0.28	20	2	8
A026	-1.98	0.91	0.92	0.48	31	2	8
B071	-2.07	1.04	0.42	0.19	43	3	7
A035	-2.13	1.03	0.62	0.28	29	3	7
B059	-2.16	1.12	0.49	0.19	27	1	9
A003	-2.16	1.12	0.25	0.13	23	2	8
A023	-2.18	1.10	0.35	0.16	21	1	9
A075	-2.23	0.88	0.90	0.55	30	2	8
A013	-2.27	0.83	0.72	0.46	27	3	7
A072	-2.35	0.83	1.12	0.95	47	3	7
B070	-2.37	0.93	1.26	0.65	19	2	8
B065	-2.44	1.12	0.61	0.22	22	1	9
B050	-2.52	1.07	0.71	0.29	21	1	9
B063	-2.65	1.91	0.10	0.03	25	0	10
A047	-2.65	1.34	0.15	0.07	16	2	8
A032	-2.79	1.63	0.09	0.04	32	1	9
A027	-2.99	0.90	0.72	0.48	18	3	7
A022	-3.14	1.00	0.60	0.32	19	1	9
B030	-3.42	2.00	0.20	0.04	23	0	10
B007	-3.45	1.88	0.07	0.03	17	0	10
B069	-3.64	1.95	0.14	0.03	12	0	10
B020	-4.07	1.37	0.14	0.07	18	1	9
B011	-4.60	1.91	0.09	0.03	16	0	10
B080	-4.83	1.93	0.12	0.03	14	0	10

B029	-5.19	1.90	0.08	0.03	13	0	10
A044	-5.54	1.95	0.14	0.03	15	0	10

Psychology 1 pinpointing PCJ P1 grade A

ID	Measure	Measure.SE	Infit	Outfit	Mark	Chosen	NotChosen
A008	2.10	0.82	0.64	0.40	2017	9	3
B004	1.82	0.74	1.18	1.03	2018	9	3
A018	1.74	0.98	1.05	1.74	2017	9	3
A015	1.19	0.74	1.12	0.95	2017	9	3
A019	0.89	0.72	0.88	0.74	2017	9	3
B039	0.87	0.77	0.52	0.40	2018	7	5
B031	0.80	0.76	0.93	0.75	2018	8	4
B082	0.73	0.67	0.97	1.03	2018	8	4
A040	0.29	0.62	1.05	1.03	2017	6	6
A013	0.26	0.69	0.97	1.29	2017	7	5
A010	0.24	0.69	1.33	1.32	2017	7	5
B006	0.15	0.63	1.05	1.05	2018	7	5
B032	0.12	0.66	1.24	1.41	2018	6	6
A004	0.11	0.66	0.94	0.90	2017	5	7
A028	0.10	0.64	0.69	0.61	2017	6	6
A009	0.10	0.68	0.67	0.64	2017	8	4
B033	-0.03	0.66	0.97	0.93	2018	5	7
B081	-0.05	0.69	0.95	0.91	2018	5	7
B042	-0.06	0.67	1.12	1.12	2018	7	5
B055	-0.13	0.66	0.99	0.91	2018	5	7
B057	-0.40	0.78	0.71	0.53	2018	4	8
A024	-0.48	0.68	1.01	1.04	2017	4	8
B083	-0.83	0.74	0.62	0.54	2018	4	8
A002	-0.91	0.67	0.79	0.76	2017	5	7
A014	-1.02	0.70	0.97	0.98	2017	3	9
B089	-1.03	0.72	0.95	1.86	2018	4	8
A020	-1.48	0.74	0.87	0.71	2017	3	9
A007	-1.50	0.83	1.38	1.24	2017	4	8
B030	-1.68	0.82	0.78	0.54	2018	4	8
B080	-1.91	0.79	1.01	1.14	2018	3	9

Psychology 1 pinpointing PCJ P1 grade E

ID	Measure	Measure.SE	Infit	Outfit	Mark	Chosen	NotChosen
B040	2.99	0.98	0.72	0.47	42	9	3
A022	2.11	0.98	0.70	0.39	41	11	1
B041	1.95	0.82	0.99	0.64	41	9	3
A023	1.83	1.02	0.72	0.36	41	11	1
A029	1.74	0.84	1.02	0.76	41	8	4

A012	1.55	0.82	0.73	0.56	41	9	3
A006	1.36	0.78	1.37	1.31	41	9	3
A003	0.98	0.88	0.94	0.99	41	8	4
B070	0.98	0.70	1.44	2.00	39	6	6
B045	0.95	0.72	1.00	0.93	38	7	5
A025	0.92	0.73	0.84	0.67	41	6	6
B037	0.56	0.81	0.27	0.23	38	7	5
B036	0.26	0.77	0.84	0.68	39	6	6
B086	-0.03	0.72	0.56	0.44	39	5	7
A027	-0.17	0.64	1.11	1.08	41	6	6
A017	-0.17	0.62	0.65	0.61	41	6	6
B091	-0.38	0.73	0.59	0.46	41	6	6
B014	-0.40	0.79	0.69	0.50	42	8	4
B090	-0.42	0.72	1.50	1.78	42	5	7
A026	-0.54	0.69	1.08	1.11	41	7	5
B028	-0.76	0.76	0.83	0.67	40	6	6
B035	-1.03	1.11	1.47	0.96	40	1	11
A011	-1.31	0.85	0.62	0.45	41	4	8
A001	-1.51	0.74	1.02	0.81	41	3	9
A005	-1.51	0.81	1.35	1.60	41	4	8
B087	-1.64	0.81	0.91	0.70	38	2	10
B001	-1.75	0.80	1.14	0.83	41	3	9
A016	-1.84	0.80	0.72	0.61	41	3	9
B085	-2.04	0.84	0.89	0.51	40	2	10
A021	-2.70	0.99	0.65	0.37	41	3	9

Psychology 1 pinpointing PCJ P2 grade A

ID	Measure	Measure.SE	Infit	Outfit	Mark	Chosen	NotChosen
B081	1.56	0.77	0.87	0.64	47	10	2
B045	1.41	0.68	0.88	0.77	51	10	3
B084	1.31	0.71	1.04	1.31	51	8	4
A002	1.20	0.74	1.11	0.85	51	9	3
A013	1.04	0.75	0.95	0.82	51	8	3
A012	0.94	0.73	0.98	1.30	51	8	4
B080	0.77	0.61	0.88	0.82	50	7	5
A001	0.56	0.73	1.42	1.51	51	8	4
A016	0.53	0.66	0.82	0.74	51	7	5
A019	0.52	0.64	1.13	1.17	51	6	7
A021	0.50	0.74	1.22	1.13	51	8	4
A015	0.49	0.64	0.69	0.64	51	6	6
B056	0.48	0.68	1.10	1.04	47	8	4
B016	0.46	0.64	1.03	1.05	51	6	6
A009	0.25	0.76	0.35	0.30	51	7	5
B028	0.23	0.67	0.96	0.92	50	6	6
B072	0.21	0.68	1.01	1.02	48	5	7

A024	0.03	0.69	0.62	0.54	51	7	5
B073	-0.45	0.70	0.90	0.90	49	3	9
A014	-0.64	0.70	0.93	0.83	51	5	7
A065	-0.66	0.65	0.62	0.59	51	5	7
A005	-0.71	0.68	1.58	1.83	51	5	7
A028	-0.73	0.73	1.18	1.07	51	5	7
B042	-0.82	0.65	0.82	0.75	47	5	7
B046	-1.05	0.69	1.20	1.32	49	4	8
B041	-1.20	0.71	0.96	0.77	50	3	9
B085	-1.33	0.82	0.92	1.24	49	4	8
B047	-1.46	0.85	1.04	1.07	48	2	9
A007	-1.63	0.85	0.79	0.48	51	3	9
B086	-1.80	0.97	0.62	0.37	48	2	10

Psychology 1 pinpointing PCJ P2 grade E

ID	Measure	Measure.SE	Infit	Outfit	Mark	Chosen	NotChosen
B023	3.99	1.88	0.06	0.03	31	12	0
B088	2.66	1.08	1.31	0.78	32	11	1
A018	2.65	0.80	0.79	0.66	31	10	2
A025	1.87	0.81	0.63	0.48	31	9	3
B045	1.83	0.95	0.27	0.17	32	9	3
A004	1.64	1.00	0.80	0.43	31	11	1
B051	1.19	0.81	1.00	0.95	30	6	6
A039	1.08	0.79	1.08	1.01	31	10	2
B061	0.77	0.68	1.06	1.03	30	6	6
A020	0.75	0.83	0.44	0.30	31	8	4
A022	0.75	0.77	1.44	1.25	31	8	4
B044	0.74	0.75	0.96	2.00	33	7	5
A017	0.67	0.75	0.70	0.53	31	8	4
B089	0.40	0.73	1.09	0.91	31	6	6
B083	0.04	0.73	1.06	0.86	33	5	7
B006	0.02	0.73	0.60	0.48	29	6	6
A023	-0.02	0.82	0.84	0.58	31	4	8
A026	-0.41	0.73	1.14	0.94	31	4	8
B008	-0.50	0.74	1.49	1.46	33	4	8
A011	-0.57	0.86	0.69	0.39	31	3	9
A008	-0.70	0.80	1.19	0.83	31	6	6
B090	-0.99	0.75	0.52	0.39	30	3	9
B043	-1.01	0.92	0.54	0.31	29	4	8
A027	-1.27	0.87	0.88	0.45	31	3	9
A010	-1.40	0.86	0.46	0.27	31	3	9
B082	-1.55	0.86	0.77	0.46	29	5	7
B049	-1.78	1.02	0.62	0.29	32	3	9
A003	-2.15	1.02	1.24	1.96	31	3	9
B050	-4.31	1.20	0.73	0.20	31	1	11

A006	-4.39	1.14	0.67	0.23	31	2	10

Psychology 2 P1 – RO

ID	Measure	Measure SE	Infit	Outfit	Mark
A034	7	1.06	0.75	0.12	61
A055	6.91	0.76	0.95	0.41	55
A042	5.78	0.62	1.17	0.58	62
A045	5.74	0.54	1.12	0.6	57
A071	5.18	0.46	0.89	0.81	63
B069	5.09	0.43	1.23	1.21	66
A051	4.89	0.46	1	0.67	64
A001	4.88	0.49	0.85	0.59	59
B072	4.86	0.54	0.81	0.38	63
B035	4.76	0.44	0.9	0.67	70
B062	4.69	0.44	0.93	0.75	60
B009	4.67	0.46	1.06	0.68	61
A030	4.14	0.53	0.96	0.63	52
A023	4.13	0.45	1.16	2.03	60
A007	4.08	0.45	1.22	1.09	53
B047	3.81	0.46	0.99	0.99	64
B080	3.71	0.51	1.13	0.74	59
A048	3.19	0.52	0.81	0.5	48
B063	3.18	0.54	0.8	0.47	58
A014	2.97	0.48	0.96	0.87	56
B044	2.61	0.52	1.24	1.22	67
B021	2.52	0.5	0.93	0.8	47
B073	2.48	0.48	0.65	0.48	57
A075	2.37	0.45	0.99	0.75	54
B065	2.28	0.45	0.78	0.62	52
B001	2.2	0.44	1.09	0.82	62
B023	1.93	0.46	1.54	1.47	55
B050	1.8	0.59	0.68	0.39	43
A063	1.42	0.42	1.01	0.8	58
B039	1.4	0.43	1.02	0.85	49
B027	1.3	0.45	1.4	1.01	50
B081	1.18	0.47	0.61	0.36	54
A031	1.14	0.45	1.41	2.2	49
A004	1.11	0.52	0.88	1.07	32
A028	1.05	0.48	0.57	0.37	51
A068	1.03	0.52	0.91	0.7	41
A032	0.95	0.41	0.9	0.74	50
B037	0.93	0.42	1.11	1.01	44
B045	0.72	0.55	1.46	1.55	53
A073	0.61	0.43	1.12	0.96	45
A072	0.56	0.4	0.75	0.6	46
B016	0.5	0.46	0.88	0.68	51

A008	0.48	0.5	1.21	1.32	42
A026	0.38	0.46	0.97	0.85	38
A022	0.37	0.43	1.01	0.78	40
A035	0.36	0.51	0.9	0.53	47
A062	0.33	0.71	0.6	0.26	29
A005	0.28	0.58	0.76	0.57	33
B071	0.23	0.51	0.83	0.46	48
A037	0.16	0.51	1.11	0.96	35
B004	0.14	0.82	0.81	0.18	56
B068	0.09	0.51	1.03	0.83	42
A036	-0.09	0.43	1.06	0.87	43
B030	-0.21	0.5	1.01	0.75	35
A029	-0.28	0.5	0.81	0.64	37
B075	-0.37	0.41	0.95	1.04	46
A047	-0.38	0.54	1.83	1.93	34
B006	-0.46	0.45	1.07	1.05	41
A059	-0.52	0.49	0.94	0.77	36
B051	-0.54	0.46	0.97	0.9	40
A080	-0.85	0.59	1.01	0.79	24
B026	-0.89	0.5	0.74	0.63	36
B003	-0.89	0.45	1.01	1.1	37
A013	-1.09	0.52	0.82	0.54	28
A039	-1.35	0.58	0.92	0.56	30
B020	-1.37	0.56	1.42	1.26	25
A057	-1.47	0.6	1.08	0.73	26
A009	-1.58	0.53	0.79	0.72	44
B055	-1.59	0.55	1.01	0.79	33
A011	-1.62	0.57	0.99	0.72	39
B057	-1.69	0.57	1.09	0.81	30
B070	-1.96	0.62	1.12	1.21	45
A050	-2.11	0.51	0.74	0.6	20
B059	-2.12	0.57	0.63	0.39	39
B012	-2.37	0.51	1.11	2.36	20
B036	-2.52	0.5	1.17	1.3	31
B007	-2.67	0.69	0.4	0.18	32
B031	-2.89	0.55	0.93	0.64	34
A081	-2.9	0.56	1.01	0.81	18
B032	-3.01	0.57	1.48	1.28	29
A020	-3.06	0.55	1.6	2.06	31
B008	-3.21	0.56	1	0.81	22
B034	-3.22	0.54	1.01	0.87	21
B022	-3.31	0.52	1.05	0.82	26
A070	-3.47	0.56	0.45	0.3	22
A012	-3.72	0.52	1	0.78	16
A006	-3.86	0.57	0.78	0.6	21
B013	-4.55	0.58	1.51	1.51	19
B014	-4.78	0.67	1	0.64	28
A016	-4.9	0.66	0.87	0.5	27

A003	-5.02	0.72	0.97	0.53	25
B048	-5.02	0.89	0.75	0.17	38
B029	-5.16	0.68	1.09	0.66	24
A069	-5.46	0.65	0.73	0.41	19
B011	-6.07	0.81	1.15	1.93	27
A065	-6.58	0.9	0.97	0.41	17
A021	-6.91	0.81	0.97	0.61	15

Psychology 2 P2 – RO

ID	Measure	Measure SE	Infit	Outfit	Mark
A059	6.31	0.57	1.11	1.28	65
B007	6.29	0.58	0.86	0.54	62
A068	5.94	0.59	1.28	1.21	54
B008	5.4	0.47	0.76	0.52	65
A057	5.28	0.46	0.78	0.56	63
A012	5.14	0.66	1	0.66	61
B011	4.98	0.44	1.02	1.02	63
B059	4.87	0.46	1.31	1.14	64
A013	4.77	0.45	0.65	0.53	58
A050	4.77	0.44	0.99	1.05	62
A032	4.38	0.44	1.16	1.04	64
B031	3.88	0.47	1.05	0.78	53
A001	3.85	0.43	1.49	1.69	57
B050	3.77	0.43	0.57	0.44	60
B001	3.7	0.41	0.81	0.65	59
B062	3.56	0.5	1.34	1.82	57
A021	3.35	0.43	1.18	1.29	60
A014	3.2	0.45	1.37	1.32	56
B009	3.15	0.42	1.03	1.44	56
B035	3.07	0.38	0.85	0.75	61
B075	2.94	0.45	0.79	0.63	55
A035	2.93	0.44	0.88	0.73	59
B023	2.91	0.49	0.91	1.37	58
A044	2.88	0.44	1.24	1.11	55
A065	2.85	0.52	0.75	0.54	53
A045	2.76	0.66	1.12	1.09	40
A037	2.68	0.43	0.84	0.58	48
B012	2.59	0.42	0.8	0.62	52
A028	2.57	0.4	0.92	0.77	50
B030	2.53	0.43	0.88	0.66	49
B020	2.52	0.43	1.01	1.21	48
B051	2.48	0.49	1.19	1.26	54
B004	2.38	0.41	1.18	1.21	50
B037	2.1	0.46	1.21	1.05	51
B055	1.91	0.43	0.87	0.74	47
A075	1.73	0.44	1.17	1.21	45

A006	1.7	0.43	0.98	0.78	46
A022	1.44	0.43	0.86	0.75	47
A069	1.32	0.43	0.76	0.58	44
A027	1.29	0.45	0.83	0.64	52
A048	0.93	0.47	1.24	1.32	49
B072	0.89	0.4	0.88	0.79	42
B065	0.8	0.45	1.09	1.08	46
A039	0.71	0.4	0.99	0.9	41
B026	0.66	0.48	1.28	1.18	35
B080	0.53	0.48	0.85	0.71	40
B036	0.49	0.4	0.88	0.78	45
B029	0.38	0.49	1.15	1.2	37
B013	0.33	0.48	1.03	1.52	30
A080	0.31	0.69	0.75	0.3	29
A036	0.29	0.4	1.24	1.85	39
B006	0.29	0.41	0.99	0.86	44
A047	0.15	0.66	0.9	0.59	51
B047	0.01	0.39	1.06	0.94	43
A011	-0.03	0.54	0.57	0.36	34
B048	-0.06	0.44	1.03	0.8	34
A063	-0.11	0.45	1.01	0.87	33
A072	-0.13	0.43	0.71	0.52	42
B021	-0.16	0.51	0.87	0.69	33
A026	-0.47	0.45	1.25	1.43	43
A081	-0.58	0.48	1.07	0.95	38
A042	-0.6	0.45	0.93	0.83	36
A029	-0.62	0.45	0.91	0.74	37
B028	-0.67	0.54	1.4	1.33	31
B057	-0.88	0.44	0.93	0.65	39
A073	-0.99	0.61	0.94	0.61	26
A071	-0.99	0.53	0.87	0.6	31
B034	-0.99	0.49	0.64	0.5	38
B027	-1.1	0.53	1.04	0.83	41
A070	-1.26	0.53	1.61	2.72	35
A034	-1.67	0.56	1.17	0.79	27
A030	-1.87	0.59	0.67	0.42	25
B070	-1.95	0.58	0.81	0.49	29
B071	-2.08	0.65	0.52	0.27	36
A016	-2.3	0.68	1.3	1.53	21
B016	-2.57	0.56	0.71	0.44	28
A031	-2.7	0.59	1.29	1.43	24
A051	-2.75	0.76	0.84	0.99	32
B073	-2.76	0.52	0.97	0.63	27
B039	-2.95	0.53	1.33	1.27	26
B042	-3	0.56	1.14	1.58	23
B069	-3.06	0.6	1	0.54	22
B063	-3.2	0.53	0.66	0.43	32
B068	-3.38	0.59	0.71	0.37	24

A004	-3.4	0.46	0.82	0.71	23
B005	-4.31	0.57	0.9	0.87	21
B044	-4.78	0.65	0.84	0.46	25
A003	-4.84	0.7	1.37	1.07	28
A062	-4.87	1.06	1.19	1.59	30
B022	-5.3	0.57	0.6	0.35	20
B045	-5.32	0.55	1.3	1.83	18
A055	-5.35	0.58	1.52	1.35	20
B032	-5.7	0.55	1.32	0.96	15
A023	-5.87	0.58	0.84	0.47	19
A007	-6.2	0.71	0.76	0.28	22
B014	-6.24	0.54	0.73	0.45	19
A005	-6.53	0.54	1.22	0.79	18
A020	-6.61	0.56	1.27	1.04	16
A008	-6.78	0.71	0.8	0.35	17
A009	-7.41	0.6	0.86	0.46	15
B003	-7.54	0.59	0.96	0.56	17

Psychology 2 P1 – PCJ

ID	Measure	Measure.SE	Infit	Outfit	Mark	Chosen	NotChosen
A014	5.66	1.93	0.12	0.03	56	10	0
A071	5.53	1.92	0.10	0.03	63	10	0
A042	5.45	1.93	0.12	0.03	62	10	0
A030	5.43	1.93	0.12	0.03	52	10	0
A034	4.92	1.91	0.09	0.03	61	10	0
A045	4.74	1.98	0.18	0.04	57	10	0
A055	4.59	1.90	0.08	0.03	55	10	0
B035	3.78	1.23	0.15	0.09	70	8	2
B080	3.64	1.47	0.14	0.06	59	9	1
B063	3.60	1.32	0.15	0.08	58	8	2
B069	3.39	1.34	0.43	0.12	66	8	2
B001	3.36	1.93	0.12	0.03	62	10	0
A051	3.11	1.14	0.44	0.16	64	9	1
B016	3.02	1.36	0.11	0.07	51	8	2
B023	2.87	1.52	0.08	0.05	55	8	2
A007	2.60	1.01	0.95	0.64	53	9	1
B021	2.52	1.56	0.13	0.05	47	9	1
B009	2.32	1.21	0.27	0.11	61	9	1
A063	2.19	1.05	1.23	0.94	58	9	1
B004	2.19	1.02	0.76	0.37	56	8	2
B062	2.08	1.55	0.08	0.05	60	7	3
A048	1.90	0.76	0.64	0.49	48	5	5
A026	1.90	1.26	0.18	0.09	38	8	2
B081	1.88	1.09	0.42	0.18	54	6	4
A001	1.86	1.12	0.68	0.24	59	9	1
B072	1.84	1.00	0.91	0.59	63	8	2
A075	1.83	1.00	0.58	0.27	54	8	2

A009	1.76	1.29	0.24	0.09	44	8	2
B045	1.68	0.94	0.45	0.24	53	7	3
B047	1.67	0.92	1.54	1.91	64	7	3
B044	1.63	0.85	0.48	0.33	67	6	4
B065	1.60	0.88	0.91	0.55	52	6	4
B050	1.59	0.75	0.82	0.72	43	7	3
A068	1.36	0.82	1.30	0.98	41	6	4
B039	1.26	0.88	0.51	0.32	49	7	3
A028	1.21	0.89	0.40	0.26	51	6	4
B070	0.98	0.83	1.02	0.89	45	6	4
B037	0.96	0.88	0.88	0.65	44	8	2
A008	0.96	0.86	0.38	0.26	42	7	3
B027	0.95	0.97	0.43	0.22	50	5	5
A032	0.94	1.00	0.58	0.27	50	7	3
B073	0.82	0.92	0.54	0.31	57	7	3
A023	0.76	0.86	0.36	0.25	60	7	3
A022	0.69	0.99	0.79	0.41	40	6	4
A037	0.44	0.99	0.23	0.15	35	6	4
A073	0.44	0.99	0.59	0.29	45	6	4
A072	0.38	0.83	1.46	1.53	46	6	4
A005	0.35	0.78	1.03	0.89	33	4	6
B071	0.34	0.86	0.59	0.36	48	7	3
A029	0.34	0.86	0.50	0.32	37	6	4
A011	0.24	0.87	0.36	0.25	39	6	4
B075	0.14	0.84	1.17	1.16	46	3	7
B059	0.12	0.96	0.33	0.19	39	5	5
A004	0.06	0.88	0.46	0.29	32	6	4
B003	-0.02	0.93	1.43	0.89	37	4	6
A062	-0.30	0.99	0.27	0.17	29	4	6
A020	-0.41	1.09	0.26	0.14	31	3	7
A059	-0.48	0.91	1.03	0.70	36	4	6
B006	-0.54	0.78	0.40	0.33	41	5	5
B032	-0.81	1.06	0.58	0.25	29	1	9
A031	-0.91	1.08	0.16	0.12	49	3	7
B030	-0.95	0.78	0.62	0.47	35	4	6
B051	-0.99	1.21	0.17	0.09	40	2	8
A036	-1.00	0.87	1.20	0.81	43	5	5
B057	-1.03	1.10	0.79	0.29	30	3	7
A070	-1.05	1.01	1.66	1.89	22	2	8
B014	-1.16	1.01	0.32	0.18	28	3	7
B007	-1.25	0.88	0.38	0.25	32	3	7
A035	-1.26	1.08	0.31	0.15	47	4	6
A013	-1.35	1.04	0.28	0.16	28	3	7
A039	-1.43	0.96	0.70	0.34	30	2	8
A047	-1.50	0.84	0.78	0.50	34	3	7
B068	-1.54	1.11	0.28	0.14	42	3	7
A006	-1.64	1.21	0.40	0.14	21	2	8
A050	-1.66	1.04	0.92	0.44	20	4	6

A057	-1.69	1.20	0.34	0.13	26	1	9
B022	-1.75	0.96	1.83	1.44	26	2	8
A080	-1.94	1.05	0.37	0.17	24	2	8
B013	-1.95	0.90	0.76	0.45	19	2	8
B036	-2.14	0.81	0.73	0.58	31	3	7
B055	-2.34	1.16	0.36	0.14	33	2	8
B008	-2.66	1.27	0.17	0.09	22	2	8
B020	-2.66	0.96	0.55	0.29	25	3	7
B012	-2.67	0.76	1.00	0.78	20	3	7
A081	-2.69	0.91	0.57	0.32	18	2	8
B011	-2.76	1.72	0.21	0.05	27	1	9
A044	-2.92	0.75	0.73	0.56	67	3	7
A003	-2.97	1.01	0.88	0.48	25	2	8
B048	-3.26	1.32	0.30	0.10	38	2	8
B026	-3.26	1.01	0.73	0.32	36	2	8
A016	-3.36	0.94	0.82	0.38	27	3	7
B029	-3.53	1.02	0.69	0.29	24	2	8
B034	-3.64	1.13	0.59	0.21	21	1	9
B031	-3.65	1.90	0.08	0.03	34	0	10
A012	-3.65	0.96	0.71	0.34	16	2	8
A027	-4.13	1.96	0.15	0.03	66	0	10
A069	-4.15	1.53	0.26	0.07	19	1	9
B028	-4.82	2.03	0.24	0.04	18	0	10
A021	-5.00	1.25	0.23	0.10	15	1	9
A065	-5.13	1.92	0.11	0.03	17	0	10
B042	-5.87	1.93	0.12	0.03	17	0	10
B005	-5.97	2.11	0.34	0.04	15	0	10

Psychology 2 P2 – PCJ

ID	Measure	Measure.SE	Infit	Outfit	Mark	Chosen	NotChosen
B059	5.08	1.88	0.06	0.03	64	10	0
A032	4.34	1.90	0.08	0.03	64	10	0
B050	4.26	1.91	0.09	0.03	60	10	0
A021	4.24	1.89	0.07	0.03	60	10	0
A057	4.23	1.90	0.09	0.03	63	10	0
B023	4.02	1.89	0.07	0.03	58	10	0
A001	3.48	1.93	0.12	0.03	57	10	0
A059	2.92	1.01	1.17	1.75	65	9	1
B031	2.67	1.00	0.92	0.63	53	9	1
A035	2.60	1.00	1.07	0.82	59	8	2
B008	2.38	0.90	0.47	0.27	65	8	2
A012	2.37	1.00	0.57	0.28	61	7	3
A045	2.34	1.03	0.74	0.37	40	9	1
A068	2.31	0.98	0.70	0.40	54	8	2
A013	2.28	1.00	0.87	0.44	58	8	2
A065	2.23	1.04	1.14	0.78	53	9	1
B007	2.20	1.04	1.31	2.12	62	9	1

B062	2.18	0.86	0.55	0.34	57	8	2
B004	2.09	0.94	0.65	0.36	50	7	3
B012	2.05	0.95	0.85	0.68	52	8	2
A014	1.99	1.09	0.24	0.13	56	8	2
B009	1.97	0.85	0.72	0.45	56	6	4
B080	1.83	0.92	0.72	0.43	40	7	3
A026	1.66	0.70	0.88	0.85	43	6	4
B037	1.60	0.91	0.62	0.33	51	7	3
A050	1.46	0.97	0.47	0.25	62	6	4
B020	1.45	0.94	1.05	0.62	48	7	3
A044	1.38	1.09	1.28	0.58	55	7	3
A028	1.36	0.68	0.84	0.83	50	5	5
B075	1.34	0.82	0.43	0.32	55	5	5
B065	1.34	0.80	0.63	0.43	46	7	3
B030	1.30	0.76	1.34	1.09	49	6	4
B011	1.26	0.94	0.36	0.22	63	8	2
B035	1.25	0.83	0.67	0.43	61	6	4
A047	1.19	0.82	0.44	0.32	51	6	4
A022	1.10	0.85	0.39	0.27	47	7	3
A048	1.08	0.76	0.98	0.92	49	5	5
A006	1.08	0.90	0.33	0.22	46	6	4
A011	0.92	0.76	1.30	0.98	34	6	4
A037	0.92	0.78	0.79	0.58	48	7	3
B051	0.91	1.03	0.23	0.14	54	6	4
B021	0.74	1.02	1.26	0.88	33	5	5
B048	0.52	0.98	1.23	0.80	34	6	4
B001	0.50	0.92	1.33	1.23	59	6	4
A075	0.48	0.80	1.10	0.80	45	4	6
B055	0.41	0.84	2.06	2.11	47	5	5
A039	0.38	0.73	0.94	1.01	41	3	7
B072	0.36	0.84	0.71	0.46	42	5	5
B028	0.28	0.87	0.80	0.54	31	5	5
A072	0.28	0.88	0.75	0.45	42	4	6
B063	0.20	0.84	0.98	0.73	32	5	5
B047	0.17	0.88	0.46	0.28	43	6	4
B026	0.15	0.76	1.22	2.16	35	5	5
B029	0.14	0.83	0.41	0.29	37	7	3
A036	0.01	0.77	0.69	0.50	39	5	5
B070	-0.03	0.88	0.40	0.26	29	6	4
B057	-0.23	0.80	0.56	0.41	39	4	6
B034	-0.34	0.76	1.02	0.82	38	5	5
A073	-0.38	0.82	0.84	0.55	26	2	8
A027	-0.41	0.98	0.35	0.20	52	5	5
A029	-0.53	0.78	0.62	0.49	37	4	6
A080	-0.71	0.87	0.47	0.30	29	3	7
B036	-0.71	0.84	0.60	0.38	45	6	4
A042	-0.77	1.02	1.61	2.24	36	4	6
B006	-0.77	0.94	0.87	0.46	44	3	7

A081	-0.86	0.84	1.06	0.74	38	4	6
B071	-0.86	0.86	1.14	0.71	36	5	5
A062	-0.97	0.83	0.68	0.47	30	2	8
B081	-1.02	1.08	0.67	0.26	66	3	7
A069	-1.09	0.95	0.79	0.43	44	6	4
A063	-1.16	0.86	0.79	0.74	33	2	8
A051	-1.23	0.86	0.73	0.45	32	3	7
A071	-1.24	0.95	0.27	0.18	31	3	7
B027	-1.28	1.21	0.39	0.14	41	1	9
B068	-1.28	0.85	0.79	0.53	24	6	4
A030	-1.33	1.33	0.11	0.07	25	3	7
A070	-1.38	0.95	0.36	0.21	35	5	5
A031	-1.55	0.83	1.75	1.78	24	3	7
A055	-1.64	0.97	1.23	1.02	20	3	7
B039	-1.67	0.92	0.50	0.30	26	3	7
B069	-1.68	0.95	0.53	0.28	22	4	6
B044	-1.75	1.00	0.56	0.26	25	3	7
A004	-1.88	0.75	0.91	0.76	23	5	5
B013	-2.00	1.09	0.97	0.40	30	1	9
B073	-2.02	1.02	0.93	0.43	27	3	7
B022	-2.13	1.33	0.22	0.09	20	2	8
B042	-2.14	0.91	0.54	0.31	23	4	6
A003	-2.24	1.01	0.73	0.38	28	1	9
A020	-2.85	0.93	1.00	0.54	16	2	8
B005	-2.94	1.22	0.26	0.11	21	1	9
A034	-3.14	1.61	0.06	0.04	27	1	9
B045	-3.21	1.16	1.57	0.76	18	1	9
B016	-3.42	0.87	0.84	0.48	28	2	8
A016	-3.51	1.16	0.73	0.23	21	1	9
A007	-3.78	1.02	0.58	0.28	22	1	9
A023	-3.86	1.43	0.13	0.06	19	1	9
B032	-3.87	1.33	0.14	0.07	15	1	9
B014	-4.10	1.87	0.04	0.03	19	0	10
A008	-4.25	1.89	0.07	0.03	17	0	10
A009	-4.34	1.88	0.06	0.03	15	0	10
A005	-5.23	1.98	0.18	0.04	18	0	10
B003	-5.53	1.93	0.12	0.03	17	0	10

English language 1 P1 PCJ

ID	Measure	Measure.SE	Infit	Outfit	Mark	Chosen	NotChosen
B21	8.42	2.06	0.28	0.02	71	19	0
A13	7.34	1.24	0.20	0.05	71	19	1
A36	6.80	1.96	0.15	0.02	64	20	0
A24	6.77	2.04	0.24	0.02	70	20	0
B04	6.10	1.28	0.17	0.04	70	18	1
B07	5.63	0.93	0.55	0.16	62	16	2
A52	5.52	1.07	0.40	0.09	59	19	1
B34	5.49	1.06	1.09	0.29	60	18	1

B55	5.37	0.91	0.72	0.21	64	19	2
A50	4.85	1.08	0.32	0.08	62	18	2
A45	4.70	1.03	1.26	0.69	66	19	1
B19	4.62	0.96	0.26	0.09	68	15	4
A12	4.40	0.93	1.12	0.28	58	19	2
A39	4.39	1.18	0.16	0.04	57	21	2
A47	4.32	0.82	1.07	0.44	65	16	3
B12	4.21	0.88	0.54	0.15	66	19	3
B35	4.17	0.76	1.19	0.56	69	15	4
B08	3.95	0.91	0.37	0.11	58	17	3
A44	3.93	0.81	0.66	0.24	63	17	3
B39	3.68	0.81	0.78	0.34	65	15	4
A49	3.61	1.24	0.19	0.05	56	17	3
B48	3.61	0.85	0.49	0.17	67	15	4
A28	3.53	0.78	0.60	0.22	68	18	3
B25	3.27	0.93	0.43	0.14	59	17	2
A54	3.09	0.73	0.73	0.33	61	15	5
A17	3.08	0.74	1.36	0.67	50	16	5
B50	2.93	0.77	0.80	0.30	61	16	4
A27	2.92	1.07	1.89	0.97	49	15	5
B45	2.90	0.82	0.76	0.27	52	16	3
B52	2.80	0.79	0.43	0.16	56	15	6
B24	2.77	0.68	0.54	0.25	48	15	7
A56	2.73	0.80	0.58	0.22	60	13	6
A19	2.72	0.76	0.68	0.28	69	16	4
B14	2.44	0.75	0.94	0.42	57	13	6
B31	2.39	0.86	0.54	0.18	63	12	8
A06	2.38	0.76	0.88	0.39	54	13	7
B29	2.26	0.82	0.46	0.18	54	16	3
A07	2.20	0.73	1.51	1.08	67	12	8
B13	2.13	0.79	0.69	0.28	41	13	7
B43	1.97	0.65	0.51	0.28	36	13	7
A35	1.95	0.67	0.98	0.53	52	13	7
B20	1.71	0.78	0.71	0.29	53	13	7
B16	1.67	0.60	0.82	0.50	55	11	9
A38	1.38	0.70	0.75	0.49	44	10	10
A46	1.37	0.66	1.19	0.68	55	11	9
B09	1.27	0.57	0.47	0.32	46	10	11
B17	1.24	0.81	0.31	0.13	44	10	11
B38	1.05	0.68	0.86	0.43	45	11	8
B53	1.02	0.59	0.46	0.28	49	12	10
A43	0.84	0.66	0.73	0.39	53	12	9
B02	0.74	0.74	0.74	0.35	50	12	7
A15	0.73	0.71	0.92	0.43	33	12	8
B32	0.70	0.72	0.88	0.42	39	10	9
A48	0.65	1.02	0.43	0.11	40	12	8

B27	0.59	0.95	0.41	0.12	51	10	9
B49	0.46	0.75	0.81	0.33	34	10	9
B47	0.32	0.78	0.54	0.21	40	11	8
A55	0.32	0.71	0.82	0.36	51	12	8
B51	0.25	0.79	1.09	0.97	42	9	10
B40	0.25	0.75	0.82	0.34	47	14	7
A33	0.18	0.70	0.59	0.28	47	9	11
A41	0.14	0.74	1.28	0.49	36	11	11
A05	-0.18	0.68	0.87	0.41	35	9	12
A32	-0.22	0.66	0.60	0.34	38	9	11
B22	-0.40	0.95	0.84	0.24	37	7	12
A20	-0.43	0.87	0.29	0.11	37	11	9
A25	-0.44	0.66	0.93	0.47	42	9	13
A22	-0.67	0.84	0.48	0.17	45	7	13
B56	-0.80	0.72	0.59	0.31	32	8	12
B26	-0.95	0.81	0.60	0.23	33	7	13
B06	-1.11	0.93	0.25	0.10	35	7	12
A53	-1.17	0.97	0.17	0.07	48	6	14
A29	-1.18	0.97	0.30	0.09	31	6	14
A03	-1.25	0.83	0.19	0.10	46	9	11
A51	-1.36	0.87	0.91	0.36	41	8	12
B11	-1.54	0.90	0.56	0.20	38	10	9
B42	-1.63	0.74	0.62	0.32	43	9	11
A02	-2.29	0.88	0.24	0.10	26	7	13
A08	-2.38	0.88	0.24	0.10	39	7	14
A18	-2.40	0.89	0.54	0.19	14	6	14
B33	-2.56	1.03	0.29	0.09	15	5	14
A34	-2.77	1.00	0.37	0.10	10	4	16
B03	-2.82	0.88	0.80	0.34	16	5	14
A16	-2.84	1.06	0.37	0.09	43	4	16
A37	-2.88	0.91	0.60	0.15	32	6	15
B44	-2.94	1.13	0.14	0.05	21	5	14
B36	-3.14	0.89	0.37	0.13	31	4	16
A10	-3.23	0.90	0.46	0.14	21	6	14
B28	-3.34	0.95	0.28	0.10	13	5	14
A31	-3.34	1.07	0.15	0.05	34	3	19
B18	-3.46	0.99	0.32	0.09	26	4	16
A23	-4.30	1.01	0.43	0.11	13	3	17
B10	-4.34	0.91	0.42	0.12	12	4	17
B46	-4.36	1.15	0.35	0.07	11	1	19
A04	-4.46	0.97	0.19	0.08	16	4	16
A11	-4.51	1.08	0.32	0.08	12	2	17
A40	-4.88	1.08	0.41	0.08	9	2	20
A26	-5.01	0.93	0.41	0.12	11	3	17
B01	-5.30	1.04	0.34	0.08	10	2	20
A42	-5.69	1.27	0.32	0.05	15	1	20

B23	-5.73	1.30	0.25	0.05	9	1	18
B30	-5.88	1.23	0.15	0.05	14	2	17
B41	-5.93	1.30	0.25	0.05	8	1	18
B15	-6.02	1.33	0.35	0.06	7	1	18
A09	-6.65	1.63	0.23	0.03	8	1	20
B37	-6.91	2.02	0.23	0.02	5	0	20
A14	-6.92	1.94	0.13	0.02	5	0	21
B54	-7.11	1.09	0.26	0.07	6	2	17
A30	-7.18	1.99	0.20	0.02	6	0	22
A01	-7.29	1.89	0.07	0.02	4	0	21
A21	-8.09	2.11	0.33	0.02	7	0	24
B05	-8.90	1.93	0.12	0.02	4	0	19

English language 1 P2 PCJ

Id	Measure	true.score.SE	infit	outfit	Mark	Chosen	NotChosen
B50	7.68	1.90	0.08	0.02	66	20	0
B46	5.65	1.10	1.04	0.22	69	18	1
B18	5.61	0.98	0.38	0.11	67	18	2
B33	5.39	1.57	0.12	0.03	65	18	1
B03	4.94	0.84	0.53	0.20	61	17	2
B29	4.73	0.87	1.62	0.69	68	17	2
A35	4.64	1.01	0.70	0.18	58	19	1
B45	4.54	1.07	0.36	0.09	48	18	1
B37	4.37	0.96	1.14	0.36	57	18	2
A08	4.36	1.23	0.15	0.04	59	18	2
B52	4.23	0.79	0.88	0.38	63	18	2
A15	4.14	0.84	0.78	0.24	66	18	2
A50	4.12	0.83	0.87	0.32	64	18	2
A34	4.10	0.74	1.16	0.76	63	16	4
B22	3.88	0.86	0.42	0.15	62	17	2
B44	3.85	0.78	0.99	0.42	58	16	3
A47	3.84	0.73	0.57	0.25	44	16	3
A19	3.84	0.77	1.34	0.88	62	17	3
A39	3.82	0.65	0.84	0.36	65	19	5
B14	3.81	0.73	0.70	0.30	59	16	4
B30	3.80	0.86	0.70	0.22	45	17	2
A27	3.75	0.77	0.56	0.22	68	17	3
A54	3.72	0.77	1.28	0.89	67	15	4
A40	3.68	0.72	1.17	0.88	48	18	3
A36	3.53	0.85	0.46	0.15	46	17	3
A17	3.17	0.68	0.67	0.34	35	15	5
A23	2.97	0.84	1.22	0.50	56	16	3
A09	2.94	0.74	1.26	0.65	50	17	4
B38	2.89	0.66	0.69	0.53	39	12	8
A49	2.89	0.68	0.87	0.50	61	14	6

A12	2.86	0.79	1.74	0.99	69	15	5
A28	2.81	0.70	0.60	0.28	60	15	5
B07	2.76	0.91	0.51	0.16	56	14	5
B48	2.71	0.75	0.74	0.68	54	15	4
B27	2.61	0.71	0.69	0.37	60	14	7
B36	2.60	0.72	0.40	0.20	64	14	6
B19	2.55	0.72	0.61	0.28	51	14	5
B09	2.26	0.73	0.43	0.20	47	13	7
A56	1.97	0.66	1.12	0.82	57	14	7
B54	1.96	0.62	0.71	0.43	55	13	6
B21	1.94	0.86	0.73	0.26	38	15	5
A10	1.94	0.79	0.50	0.18	55	14	7
B51	1.88	0.69	0.86	0.47	53	12	7
A29	1.67	0.75	0.61	0.24	54	13	8
A55	1.65	0.69	0.88	0.43	53	14	7
B55	1.61	0.67	0.95	0.56	50	11	8
B56	1.59	0.66	0.43	0.24	52	13	7
B17	1.39	0.73	0.85	0.80	43	12	7
B11	1.35	0.70	1.02	0.85	40	11	8
B10	1.18	0.75	0.83	0.77	34	10	10
A33	1.14	0.85	1.39	0.59	39	10	10
B02	0.97	0.73	0.50	0.23	46	12	8
A24	0.88	0.71	0.81	0.37	51	7	13
A48	0.83	0.82	0.43	0.15	42	11	11
B23	0.83	0.76	0.75	0.33	41	10	10
A20	0.83	0.78	0.62	0.24	40	12	8
A44	0.75	0.85	0.36	0.15	41	8	12
B26	0.75	0.79	1.08	0.64	44	10	10
A18	0.59	0.76	0.62	0.23	37	12	11
A43	0.53	0.67	0.56	0.29	45	9	11
A41	0.40	0.70	1.13	0.85	33	8	13
A14	0.27	0.75	0.55	0.23	47	13	7
B05	0.25	0.85	0.40	0.15	32	11	8
B53	0.22	0.84	0.32	0.13	37	11	9
B41	0.08	0.82	0.93	0.38	42	9	11
B42	-0.05	0.74	1.01	0.49	35	10	10
A53	-0.48	0.78	0.28	0.12	43	8	14
B40	-0.56	0.86	0.25	0.11	30	10	10
B49	-0.71	0.86	0.72	0.27	49	7	12
A46	-0.91	0.99	0.27	0.08	34	8	12
A42	-1.21	0.93	0.74	0.24	49	8	14
B28	-1.27	0.78	0.39	0.16	33	7	13
B32	-1.30	1.09	0.26	0.07	36	4	15
B35	-1.38	0.95	0.31	0.10	12	7	12
A26	-1.42	0.88	0.28	0.11	29	9	11
A30	-1.47	0.81	0.79	0.33	32	5	16

A06	-1.59	0.84	0.35	0.13	31	8	12
B47	-1.64	0.84	0.37	0.14	31	7	13
B06	-1.64	1.04	0.18	0.07	20	4	15
A13	-1.77	0.79	0.62	0.27	52	6	14
A38	-1.86	1.05	0.14	0.06	30	5	14
B15	-2.47	0.95	0.31	0.09	29	4	18
A05	-2.63	0.75	0.31	0.13	24	5	19
A37	-2.97	1.20	0.17	0.05	38	2	19
A11	-2.98	0.75	0.59	0.24	20	6	14
A51	-3.18	0.67	0.43	0.24	14	6	14
A07	-3.27	0.85	0.55	0.17	15	6	14
A22	-3.44	0.82	0.51	0.18	13	6	14
B43	-3.48	1.02	0.77	0.17	24	4	16
A45	-3.66	1.04	0.64	0.13	36	2	18
B24	-3.92	1.00	0.35	0.09	13	3	17
B25	-3.99	1.38	0.15	0.04	9	2	17
A01	-4.13	0.80	0.56	0.17	12	4	20
B31	-4.22	1.88	0.12	0.02	16	1	18
B12	-4.27	0.92	0.42	0.13	15	3	16
A02	-4.52	1.04	0.74	0.16	16	3	17
B16	-4.60	1.09	0.56	0.11	10	1	18
B01	-4.78	1.06	0.30	0.08	11	4	15
A52	-4.81	1.20	0.20	0.05	9	4	16
A21	-4.82	1.19	0.23	0.05	10	3	17
B20	-5.60	1.08	0.31	0.08	6	2	17
B08	-5.90	1.35	0.23	0.04	8	1	19
A32	-5.95	0.97	0.21	0.08	6	3	18
A31	-5.98	1.21	0.34	0.06	11	2	20
A04	-6.26	1.68	0.15	0.02	8	1	22
A25	-6.75	1.51	0.11	0.03	5	1	19
B39	-6.91	2.12	0.34	0.02	7	0	19
B34	-7.28	1.15	0.20	0.06	14	2	17
A03	-7.73	1.91	0.10	0.02	4	0	21
A16	-7.97	1.92	0.11	0.02	7	0	20
B13	-8.85	1.97	0.16	0.02	4	0	19
B04	-8.91	1.93	0.12	0.02	5	0	20

English language 2 P1 RO

Id	Measure	true.score.SE	infit	outfit	Mark	Chosen	NotChosen
A42	6.43	0.64	0.72	0.38	56	31	3
A43	6.32	0.52	1.12	0.77	66	30	5
A56	6.15	0.45	0.89	0.69	67	27	8
A53	5.35	0.41	0.94	0.87	62	25	10
A55	5.20	0.37	0.73	0.64	72	23	17
A47	4.72	0.39	0.95	0.84	70	16	19
A45	4.52	0.41	1.25	1.18	64	26	14
A57	4.51	0.39	1.31	1.32	68	17	18
A14	4.47	0.39	0.76	0.69	58	17	18
A50	4.39	0.40	0.91	0.71	61	18	17
A23	4.05	0.50	1.08	0.78	48	26	8
A18	3.89	0.42	1.27	1.34	52	16	19
A12	3.89	0.38	0.83	0.69	65	14	26
A58	3.82	0.41	1.07	1.10	63	12	23
A20	3.60	0.42	0.85	0.81	53	16	19
A54	3.25	0.44	1.00	1.10	60	14	26
A09	3.00	0.45	0.58	0.38	54	18	17
A31	2.92	0.44	0.75	0.55	46	20	14
A24	2.55	0.50	0.91	0.62	51	9	26
A17	2.39	0.48	0.81	0.74	40	10	24
A21	2.34	0.44	1.08	1.18	44	18	17
A40	1.66	0.43	0.76	0.63	41	16	18
A28	1.42	0.52	1.50	1.06	45	14	21
A27	1.33	0.49	1.08	0.72	34	27	8
A41	1.31	0.50	0.69	0.50	57	8	27
A06	0.94	0.58	1.16	0.76	50	6	29
A30	0.93	0.60	0.35	0.16	55	5	29
A25	0.85	0.45	0.92	0.81	49	14	19
A38	0.78	0.47	1.12	0.68	39	27	12
A19	0.49	0.45	1.05	0.99	32	20	14
A22	0.17	0.47	0.93	0.66	30	23	17
A02	0.14	0.45	1.12	1.15	38	14	21
A10	-0.01	0.50	0.79	0.43	35	24	11
A33	-0.08	0.49	0.83	0.48	47	8	32
A16	-0.29	0.59	0.65	0.35	36	11	24
A32	-0.34	0.50	1.47	1.15	37	14	21
A04	-0.38	0.48	0.63	0.36	19	20	15
A36	-0.63	0.46	0.96	0.84	33	15	20
A08	-1.14	0.51	0.95	0.57	29	18	15
A26	-1.23	0.58	0.88	0.45	31	12	23
A46	-1.79	0.52	1.24	0.55	25	26	14
A44	-1.93	0.53	0.94	0.56	27	21	19
A05	-2.44	0.70	0.58	0.16	16	20	15
A29	-2.69	0.62	0.81	0.26	28	11	29

A63	-4.46	0.76	0.78	0.33	14	13	20
A37	-4.76	0.84	0.35	0.07	22	6	34
A60	-5.93	0.74	0.69	0.17	12	9	25
A03	-6.85	0.70	0.42	0.12	15	22	16
A61	-7.52	0.72	0.68	0.28	10	19	16
A62	-9.36	0.53	0.93	0.49	11	13	22
A48	-9.41	0.53	1.11	0.65	7	12	23
A35	-9.41	0.57	0.87	0.40	8	13	22
A39	-9.43	0.57	1.04	0.56	9	9	26
A11	-10.02	0.49	0.85	0.46	5	10	25
A01	-10.73	0.51	0.92	0.52	6	6	27
A13	-14.41	1.85	0.02	0.01	4	0	40
B27	6.86	0.56	0.66	0.31	70	31	4
B44	6.49	0.48	1.42	1.45	68	32	6
B48	6.32	0.53	1.02	0.95	61	31	4
B02	5.97	0.42	0.83	0.90	65	30	9
B08	5.62	0.43	1.01	0.78	72	24	11
B07	5.34	0.41	0.85	0.70	54	24	11
B52	4.89	0.38	1.03	0.89	66	24	16
B35	4.81	0.41	1.44	1.57	64	19	15
B54	4.70	0.47	0.81	0.64	55	27	8
B33	4.68	0.38	0.87	0.86	63	18	17
B30	4.65	0.43	0.92	0.70	58	22	13
B10	4.42	0.40	1.03	1.09	62	18	17
B25	4.42	0.46	0.97	0.87	57	24	11
B51	4.41	0.39	1.10	1.15	52	19	16
B28	4.36	0.43	1.04	0.88	53	18	17
B40	4.05	0.43	0.89	0.89	67	17	18
B06	3.41	0.44	0.64	0.46	56	19	16
B36	3.22	0.46	1.23	1.07	51	16	19
B32	3.11	0.45	0.90	0.74	50	11	23
B50	3.08	0.48	1.08	0.88	48	25	10
B43	2.92	0.54	1.22	0.84	47	29	6
B46	2.72	0.49	1.05	0.88	60	9	25
B09	2.61	0.44	0.82	0.73	46	19	16
B17	2.47	0.43	1.08	0.85	49	19	21
B38	2.46	0.49	0.96	0.76	39	23	11
B03	2.13	0.50	0.81	0.45	32	33	6
B14	1.91	0.52	1.01	0.85	33	29	6
B01	1.90	0.60	0.96	0.37	30	36	4
B19	1.87	0.50	1.86	1.92	37	28	7
B55	1.68	0.47	0.83	0.55	38	25	10
B20	1.67	0.53	0.75	0.32	31	33	7
B15	1.46	0.41	0.87	0.79	40	15	20
B39	1.17	0.43	0.87	0.78	45	17	18
B49	0.93	0.44	1.00	0.86	44	15	20

B56	0.43	0.50	1.12	0.71	35	24	11
B37	-0.31	0.51	1.36	1.00	22	21	14
B41	-0.34	0.61	0.94	0.50	36	5	30
B23	-0.41	0.42	0.93	0.77	34	18	22
B45	-0.72	0.49	0.81	0.45	29	29	11
B05	-1.69	0.63	1.04	0.54	16	18	15
B12	-2.31	0.84	1.69	0.54	10	30	4
B34	-2.41	1.05	0.27	0.05	41	1	31
B26	-2.44	0.58	0.98	0.43	27	26	14
B13	-2.52	0.59	0.72	0.32	28	19	21
B31	-2.69	0.63	0.74	0.28	25	17	23
B21	-3.07	0.76	0.32	0.10	19	13	21
B42	-5.40	0.76	0.70	0.16	15	14	21
B22	-5.72	0.77	0.63	0.14	8	11	24
B29	-5.77	0.74	0.48	0.12	12	4	31
B47	-7.37	0.65	0.74	0.38	7	23	12
B11	-9.75	0.65	1.07	0.34	14	7	28
B24	-10.19	0.47	0.70	0.39	11	11	24
B16	-10.26	0.48	1.18	0.65	6	9	26
B18	-10.53	0.56	0.92	0.35	9	6	34
B04	-10.84	0.56	0.82	0.39	5	8	27
B53	-10.93	0.58	1.05	0.47	4	6	29

English language 2 P2 RO

Id	Measure	true.score	SE	infit	outfit	Mark	Chosen	NotChosen
B05	8.06	1.88	0.06	0.01	56	2019	35	0
A37	7.38	0.94	0.58	0.12	69	2018	34	1
A44	6.25	0.66	1.19	0.46	72	2018	36	3
B16	6.12	0.76	0.45	0.15	59	2019	31	3
A55	4.93	0.52	1.40	0.97	71	2018	24	11
B03	4.64	0.47	1.06	0.79	71	2019	26	9
B56	4.60	0.40	1.03	0.94	66	2019	29	10
A59	4.58	0.47	0.89	0.59	63	2018	27	8
A17	4.55	0.41	0.90	0.66	65	2018	29	11
A63	4.22	0.40	0.84	0.63	66	2018	23	12
B28	4.09	0.41	1.06	1.04	69	2019	27	13
B38	3.94	0.42	1.24	1.56	64	2019	22	13
B24	3.89	0.36	0.97	0.88	65	2019	22	18
B22	3.77	0.47	0.59	0.32	49	2019	37	8
A23	3.72	0.46	0.91	0.60	55	2018	25	10
B48	3.60	0.48	0.84	0.51	55	2019	24	11
A32	3.58	0.43	0.85	0.59	64	2018	23	16
B15	3.57	0.43	1.24	1.00	62	2019	21	14
A47	3.49	0.51	1.04	0.68	57	2018	28	7
B11	3.39	0.45	0.66	0.70	51	2019	20	15
B26	3.39	0.41	1.04	0.89	63	2019	19	16

B27	3.37	0.42	1.19	1.02	67	2019	20	15
A35	3.33	0.45	0.92	0.67	67	2018	16	19
A42	3.30	0.41	0.83	0.65	52	2018	19	16
A21	3.18	0.39	1.04	1.01	58	2018	18	17
B19	2.88	0.95	0.94	0.41	30	2019	39	1
B30	2.85	0.43	1.33	1.16	72	2019	13	22
A20	2.66	0.43	0.94	0.66	56	2018	20	15
A48	2.60	0.42	0.70	0.54	53	2018	16	19
B55	2.45	0.46	1.13	1.04	54	2019	16	18
A43	2.38	0.43	1.01	0.84	61	2018	13	21
A50	2.32	0.45	0.82	0.61	50	2018	17	18
A22	2.23	0.47	1.77	1.73	46	2018	21	14
B06	2.13	0.45	1.15	0.90	58	2019	11	24
A34	2.07	0.45	0.90	0.65	51	2018	16	19
B54	2.00	0.42	0.68	0.51	46	2019	20	15
B50	1.84	0.51	0.62	0.30	32	2019	34	6
B31	1.74	0.50	0.71	0.44	61	2019	11	24
A25	1.60	0.45	0.81	0.59	45	2018	17	18
A14	1.59	0.37	1.05	0.87	47	2018	22	18
B01	1.55	0.40	0.86	0.70	47	2019	25	10
B35	1.49	0.49	0.63	0.37	33	2019	28	6
A49	1.39	0.46	1.59	1.70	54	2018	14	21
B07	1.33	0.48	1.03	0.95	37	2019	26	9
B40	1.25	0.46	1.13	0.83	38	2019	26	8
A45	1.23	0.47	0.88	0.76	59	2018	8	32
B45	1.21	0.41	0.96	1.08	48	2019	19	16
A39	1.20	0.40	1.16	0.93	49	2018	19	16
A56	1.17	0.51	0.91	0.88	62	2018	7	27
B49	1.12	0.40	1.07	0.86	44	2019	17	18
B42	1.06	0.43	1.01	0.76	36	2019	16	19
A13	0.79	0.45	1.16	1.07	44	2018	13	22
A08	0.77	0.45	1.04	0.61	39	2018	30	10
B17	0.71	0.40	1.16	1.08	45	2019	18	17
A26	0.67	0.41	0.84	0.64	37	2018	21	14
B21	0.66	0.45	0.92	0.71	57	2019	9	26
B52	0.33	0.49	1.19	0.83	35	2019	25	10
B14	0.31	0.46	0.87	0.64	39	2019	15	20
A04	0.28	0.47	1.34	1.07	31	2018	22	13
A07	0.26	0.44	0.91	0.67	34	2018	24	11
B12	0.21	0.59	0.97	0.52	52	2019	5	30
A31	0.19	0.43	1.05	0.93	41	2018	12	22
A38	0.18	0.60	0.65	0.27	40	2018	4	31
A36	0.17	0.44	1.31	1.38	22	2018	25	15
A24	0.12	0.49	0.93	0.69	35	2018	25	10
B39	0.12	0.59	0.63	0.24	53	2019	4	31
B43	0.11	0.39	0.74	0.55	34	2019	24	16
B13	0.05	0.48	1.09	0.82	41	2019	8	27
A09	0.02	0.45	1.00	0.90	48	2018	11	24

A12	-0.29	0.44	0.88	0.56	30	2018	21	19
A06	-0.66	0.44	0.68	0.50	33	2018	16	19
A05	-0.74	0.48	1.48	1.25	32	2018	16	19
A40	-0.76	0.57	1.05	0.44	25	2018	32	8
A02	-0.82	0.43	1.07	0.72	26	2018	25	15
B18	-0.97	0.73	0.71	0.23	50	2019	2	33
B20	-1.15	0.52	0.90	0.48	22	2019	19	16
A41	-1.30	0.58	0.89	0.53	36	2018	9	26
A16	-1.31	0.74	0.39	0.11	15	2018	32	3
B08	-1.34	0.54	0.75	0.31	26	2019	28	11
A33	-1.42	0.51	0.88	0.86	38	2018	7	28
B41	-1.51	0.63	0.70	0.58	40	2019	3	32
A46	-1.60	0.47	0.81	0.46	29	2018	17	18
B02	-1.79	0.51	1.06	0.58	25	2019	19	21
B53	-1.81	0.52	1.01	0.74	31	2019	16	22
B37	-1.82	0.43	0.95	0.58	29	2019	23	17
B36	-2.16	0.50	0.84	0.58	28	2019	20	20
B46	-2.42	0.57	1.58	1.29	19	2019	14	19
A28	-2.53	0.60	0.50	0.19	28	2018	10	30
B34	-3.09	0.60	1.32	1.17	16	2019	15	20
A54	-3.55	0.56	0.96	0.65	11	2018	22	13
B44	-3.81	0.55	1.26	1.09	11	2019	26	9
B33	-4.20	0.55	0.44	0.22	14	2019	15	20
A57	-4.63	0.50	1.01	0.55	8	2018	17	18
B32	-4.80	0.66	0.49	0.18	15	2019	9	21
A27	-4.98	0.52	0.84	0.41	16	2018	12	27
A53	-5.05	0.70	0.72	0.20	19	2018	5	30
A58	-5.12	0.48	1.24	0.81	10	2018	17	18
A19	-5.16	0.55	1.51	1.18	14	2018	11	24
A52	-5.21	0.54	0.82	0.39	9	2018	12	23
B25	-5.30	0.44	0.66	0.44	10	2019	17	18
A15	-5.81	0.43	1.53	1.55	4	2018	13	26
B10	-5.85	0.54	0.65	0.27	9	2019	8	32
B23	-5.91	0.71	0.62	0.16	12	2019	3	32
B09	-5.95	0.46	0.84	0.51	7	2019	13	22
A51	-6.06	0.58	0.70	0.29	12	2018	7	28
A11	-6.09	0.47	0.82	0.51	7	2018	12	22
A60	-6.27	0.47	1.19	1.00	5	2018	11	24
B04	-6.52	0.67	1.06	0.36	8	2019	3	32
B47	-6.85	0.47	0.93	0.68	5	2019	7	28
A10	-7.32	0.58	0.58	0.40	6	2018	8	27
B51	-7.66	0.63	1.13	0.69	4	2019	3	32
B29	-10.65	1.85	0.02	0.01	6	2019	0	35

English language 3 P1 PCJ

Id	Measure	true.score.SE	infit	outfit	Mark	Chosen	NotChosen
B47	8.03	1.87	0.04	0.01	52	25	0

A21	6.19	0.99	0.60	0.13	56	25	1
A49	6.01	1.87	0.05	0.01	52	26	0
B12	5.43	0.82	1.54	0.59	56	23	3
B48	5.31	0.88	0.62	0.15	54	23	2
A47	5.26	1.02	0.33	0.07	61	24	2
B49	5.14	1.02	0.54	0.10	59	24	1
B08	4.83	0.79	0.76	0.37	45	23	3
B57	4.76	0.88	0.94	0.24	51	23	3
A52	4.72	0.66	0.85	0.36	44	22	5
B46	4.30	0.80	0.45	0.14	55	23	3
B55	4.11	0.68	0.44	0.18	61	22	4
B28	3.98	0.88	1.07	0.51	53	25	2
A45	3.94	0.72	1.51	0.57	51	23	4
A30	3.73	0.75	1.26	0.48	47	24	4
B19	3.70	0.75	0.97	0.40	48	22	4
A10	3.61	0.71	1.05	0.40	48	21	5
A17	3.60	0.66	1.08	0.50	54	20	6
B04	3.58	0.72	0.51	0.18	42	22	4
B63	3.52	0.68	0.61	0.25	47	22	5
B05	3.27	0.61	0.72	0.34	49	18	8
B21	3.18	0.67	0.59	0.22	57	21	6
A12	3.02	0.67	1.30	0.76	45	20	6
A25	2.86	0.64	1.10	0.56	46	20	9
A36	2.78	0.61	0.67	0.33	57	19	7
B15	2.72	0.73	0.93	0.34	41	17	9
B24	2.71	0.64	0.58	0.29	38	17	8
B02	2.70	0.56	0.81	0.44	32	17	9
A51	2.61	0.78	0.73	0.23	36	21	6
A28	2.54	0.65	1.55	0.82	53	18	8
A44	2.43	0.66	0.70	0.28	59	16	10
A31	2.22	0.57	1.37	1.03	55	16	10
B23	2.21	0.63	0.38	0.19	43	18	8
A13	2.12	0.62	0.93	0.48	42	16	10
A43	2.02	0.73	0.51	0.18	43	18	8
A50	1.98	0.64	0.62	0.26	35	20	7
B06	1.86	0.64	1.18	0.63	44	16	10
A48	1.59	0.58	0.66	0.63	41	14	12
B61	1.25	0.72	0.31	0.14	33	17	8
A29	0.86	0.68	0.99	0.79	49	13	13
A03	0.76	0.62	0.98	0.56	33	11	15
A61	0.76	0.71	0.81	0.53	19	11	15
A26	0.58	0.65	1.07	0.62	25	14	11
B30	0.53	0.74	1.09	0.58	46	11	15
A08	0.42	0.71	0.42	0.17	27	13	13
A01	0.34	0.67	0.99	0.54	34	12	14
A14	0.23	0.68	1.05	0.56	32	11	15

A23	0.22	0.64	0.80	0.38	38	11	15
A38	0.21	0.68	0.59	0.22	29	12	15
B34	0.21	0.67	0.40	0.19	34	14	11
B62	0.20	0.66	0.47	0.19	36	12	16
B14	-0.39	0.71	0.63	0.24	28	10	16
A58	-0.52	0.68	0.71	0.33	28	8	18
A57	-0.78	0.75	0.75	0.26	31	13	14
A62	-1.04	0.66	0.57	0.26	30	10	18
B45	-1.07	0.74	0.35	0.13	35	12	14
A33	-1.37	0.63	0.92	0.54	23	8	18
B44	-1.38	0.83	1.10	0.35	25	7	19
A42	-1.55	0.69	0.57	0.21	24	13	13
B13	-1.56	0.60	0.69	0.32	27	11	16
A19	-1.57	0.81	0.76	0.19	15	10	17
B36	-1.74	0.68	0.70	0.27	26	10	15
B01	-1.78	0.70	1.09	0.77	31	8	18
B59	-1.79	0.94	0.37	0.08	23	7	20
B54	-1.81	0.81	0.59	0.17	24	7	18
B52	-1.91	0.80	0.77	0.28	30	8	17
B40	-2.01	0.74	0.66	0.22	29	8	17
A05	-2.02	0.81	0.65	0.17	26	7	19
B39	-2.88	0.75	0.25	0.11	9	7	19
A63	-3.46	0.94	0.26	0.07	10	5	21
A40	-3.79	0.83	0.60	0.20	11	5	21
A60	-4.09	0.85	0.79	0.45	6	8	18
A04	-4.34	0.71	0.76	0.25	8	7	22
A15	-4.48	0.86	0.66	0.16	5	3	23
B07	-4.59	0.79	0.50	0.14	8	4	22
B31	-4.66	0.75	0.61	0.18	19	6	21
B25	-4.80	0.70	0.51	0.18	15	4	23
B50	-4.98	0.82	0.85	0.27	6	4	21
B42	-5.28	1.03	0.29	0.06	7	3	23
A59	-5.43	0.99	1.11	0.22	3	2	24
A18	-5.53	0.87	0.54	0.14	7	4	22
B29	-5.69	1.02	0.36	0.07	10	2	23
A32	-5.70	0.98	0.38	0.08	9	2	24
B20	-5.96	1.06	0.74	0.12	3	1	25
B41	-5.99	1.19	0.94	0.14	5	3	21
B17	-6.38	1.05	1.30	0.41	4	1	27
A54	-6.81	0.85	0.55	0.14	4	2	27
B10	-7.87	1.01	0.22	0.06	2	2	24
A11	-8.93	2.07	0.29	0.02	2	0	26
B18	-9.21	1.98	0.18	0.01	11	0	26

English language 3 P2 PCJ

Id	Measure	true.score.SE	infit	outfit	Mark	Chosen	NotChosen
A15	6.44	1.01	1.00	0.24	80	26	1
A10	5.77	0.87	0.72	0.17	67	24	2
B67	5.68	0.81	0.77	0.23	76	24	2
A63	5.65	0.84	0.57	0.15	81	24	2
B07	5.42	0.70	1.27	0.73	74	23	3
A81	5.26	1.04	1.01	0.23	82	24	1
B14	5.23	0.79	0.80	0.32	84	24	2
A53	5.22	0.78	1.06	0.47	84	23	3
A71	5.15	0.75	0.60	0.18	86	23	3
B18	5.13	1.00	0.57	0.11	64	25	1
B53	5.08	0.86	0.36	0.11	78	22	2
A16	5.05	0.76	1.06	0.63	71	23	3
A73	5.00	0.77	0.68	0.20	73	23	3
A93	4.84	0.81	1.21	0.45	79	21	3
B08	4.57	0.93	1.08	0.27	82	23	2
B57	4.56	0.91	0.97	0.24	80	23	2
B37	4.30	0.72	0.96	0.45	81	22	3
A70	4.27	0.68	0.61	0.24	62	22	4
B55	4.24	0.80	0.79	0.29	68	22	3
A45	4.09	0.68	0.50	0.20	69	21	5
B41	4.07	0.74	0.33	0.13	79	22	4
A94	3.93	0.73	0.43	0.16	77	21	4
A40	3.90	0.71	0.86	0.36	65	22	3
B05	3.87	0.76	1.48	0.81	59	23	4
A49	3.72	0.79	1.04	0.43	76	20	6
B49	3.69	0.57	1.13	0.70	75	16	9
B02	3.58	0.72	0.99	0.44	58	21	4
B58	3.43	0.76	0.97	0.37	86	19	6
A59	3.34	0.73	0.75	0.25	58	21	5
B39	3.22	0.63	0.94	0.67	69	18	8
B46	3.14	0.86	1.38	0.55	71	21	4
B11	3.12	0.61	0.48	0.23	65	20	7
B66	3.11	0.67	0.93	0.53	73	23	5
A32	3.03	0.78	1.61	0.89	72	21	5
A83	2.90	0.64	0.41	0.20	63	19	6
A82	2.80	0.79	0.42	0.14	68	20	6
A95	2.79	0.77	1.45	0.64	75	18	7
B23	2.69	0.70	0.30	0.12	70	21	8
A47	2.62	0.66	0.94	0.41	74	17	8
A14	2.60	0.56	0.91	0.81	70	16	10
A17	2.55	0.83	0.34	0.11	61	21	5
A58	2.49	0.66	0.65	0.32	46	17	9
A43	2.41	0.60	0.80	0.41	56	17	10
A12	2.31	0.72	0.69	0.30	50	17	9

A60	2.12	0.63	0.89	0.43	53	19	8
B10	1.92	0.61	1.20	0.65	63	18	8
B36	1.90	0.64	0.94	0.41	77	22	7
A28	1.90	0.73	0.64	0.21	78	19	8
A38	1.87	0.62	0.76	0.37	55	17	9
A26	1.70	0.62	0.83	0.42	37	16	10
B27	1.59	0.65	0.68	0.40	72	14	11
B16	1.49	0.72	0.69	0.25	57	17	9
B50	1.34	0.70	0.66	0.30	61	17	8
A20	1.31	0.56	0.98	0.89	64	15	11
B56	1.26	0.70	0.92	0.39	67	13	13
A08	1.26	0.64	0.94	0.47	57	18	9
A87	1.25	0.67	0.57	0.24	51	15	12
B34	1.19	0.61	1.20	0.80	39	16	9
B44	0.94	0.64	1.02	0.70	53	13	12
B54	0.79	0.73	0.85	0.35	60	14	11
B20	0.75	0.62	0.76	0.37	49	15	11
A23	0.71	0.58	0.81	0.42	36	12	14
B04	0.66	0.68	1.11	0.92	56	16	11
A35	0.58	0.61	0.41	0.22	48	12	14
B17	0.54	0.61	1.24	0.64	55	15	12
B63	0.49	0.63	0.77	0.48	62	17	8
A67	0.40	0.63	0.98	0.51	59	13	13
A34	0.23	0.65	0.60	0.27	42	14	12
B62	0.22	0.72	0.79	0.35	47	15	10
B19	0.15	0.73	0.68	0.25	43	12	13
B45	0.05	0.65	1.17	0.56	52	10	15
A76	-0.07	0.66	0.38	0.18	43	13	13
A84	-0.07	0.76	0.43	0.15	52	14	11
B48	-0.11	0.67	1.03	1.11	50	13	13
B31	-0.17	0.66	0.76	0.39	36	11	14
B60	-0.18	0.72	0.42	0.17	48	10	15
B15	-0.28	0.72	0.72	0.32	42	13	12
B26	-0.32	0.65	1.04	0.44	45	11	17
A33	-0.32	0.73	0.50	0.19	39	15	10
A85	-0.38	0.61	1.19	0.64	41	14	15
A44	-0.65	0.63	0.45	0.22	47	10	16
A11	-0.66	0.70	0.87	0.42	44	9	18
B03	-0.67	0.67	0.71	0.30	51	9	16
A25	-0.97	0.75	0.89	0.32	45	8	18
B33	-0.99	0.70	1.20	0.56	44	8	17
A02	-1.18	0.77	0.51	0.16	60	12	15
A88	-1.36	0.64	1.09	0.62	35	11	14
B25	-1.43	0.66	0.54	0.26	40	7	19
A22	-1.66	0.73	0.94	0.56	40	7	19
B29	-1.68	0.63	0.52	0.23	32	7	20

B40	-1.71	0.70	0.82	0.71	37	11	16
B47	-1.81	0.69	0.88	0.49	41	8	17
B35	-2.03	0.78	0.54	0.17	38	7	18
B42	-2.18	0.88	0.29	0.09	35	7	18
A64	-2.23	0.81	0.31	0.10	29	9	18
B43	-2.27	0.78	1.19	0.53	25	7	19
A37	-2.34	0.90	0.34	0.09	38	8	18
A66	-2.35	0.76	0.53	0.20	15	6	20
A91	-2.40	0.69	0.74	0.31	49	8	20
A24	-2.42	0.92	0.67	0.17	12	3	23
B65	-2.43	0.75	0.34	0.13	16	10	15
A74	-2.54	0.70	0.40	0.16	21	8	19
A09	-2.54	0.72	0.83	0.41	17	6	20
A19	-2.74	0.79	0.53	0.16	16	5	21
B61	-2.85	0.77	0.43	0.15	21	5	21
B06	-3.09	1.00	0.25	0.06	29	3	22
A18	-3.10	0.81	1.09	0.94	25	4	22
A72	-3.30	0.71	0.75	0.29	14	6	20
B28	-3.73	0.73	0.30	0.12	13	5	22
A86	-3.77	0.84	1.54	0.62	6	3	22
B59	-4.00	1.10	0.28	0.06	15	1	24
A01	-4.12	0.92	0.29	0.07	32	4	24
B21	-4.13	0.79	0.85	0.28	17	5	19
B51	-4.22	0.97	0.35	0.08	11	4	22
A51	-4.25	0.75	0.37	0.13	8	5	21
B52	-4.43	0.81	1.40	0.82	5	3	23
B12	-4.44	0.72	0.64	0.23	8	6	19
B30	-4.56	0.81	0.67	0.20	10	4	21
B32	-4.59	1.03	0.89	0.15	14	3	23
A54	-5.03	0.88	0.59	0.14	11	3	23
B22	-5.03	1.10	0.17	0.05	12	2	23
A13	-5.07	0.89	0.76	0.16	9	4	23
A77	-5.51	1.01	0.53	0.10	10	1	25
B38	-5.91	0.99	0.22	0.06	46	5	20
A89	-6.21	1.08	0.27	0.06	13	2	23
B24	-6.31	1.90	0.09	0.01	4	0	25
B09	-6.44	1.91	0.09	0.01	7	0	25
A41	-6.48	1.31	0.14	0.03	7	1	26
B64	-7.03	1.02	0.33	0.07	9	2	22
B13	-7.68	1.61	0.35	0.03	6	1	27
A79	-7.92	1.41	0.15	0.03	5	1	25
A46	-8.28	1.90	0.08	0.01	4	0	26
A48	-8.62	1.94	0.13	0.01	3	0	26
B01	-9.66	1.95	0.14	0.01	3	0	27

English language 4 P1 – RO

Id	Measure	true.score.SE	infit	outfit	Mark	Chosen	NotChosen
A044	7.88	0.94	0.72	0.25	62	24	1
B011	7.10	0.98	0.92	0.31	56	23	1
A050	6.87	0.83	0.68	0.34	59	18	2
B010	6.26	0.61	1.08	1.20	57	18	5
A068	5.39	0.73	0.72	0.31	43	22	3
A048	5.32	0.52	1.06	0.88	60	16	9
B015	5.29	0.57	1.04	0.79	58	19	6
B024	5.28	0.61	0.92	0.71	53	21	4
A073	5.16	0.54	0.79	0.51	61	17	8
B013	5.09	0.57	1.09	0.92	62	17	8
B016	4.92	0.54	1.03	0.89	41	13	7
A041	4.83	0.52	1.20	0.93	53	18	7
A056	4.69	0.65	0.91	0.56	44	16	4
B028	4.49	0.54	0.83	0.59	59	13	12
A030	4.31	0.51	0.47	0.34	51	15	10
B030	4.30	0.53	0.96	0.92	61	13	12
A060	4.28	0.52	0.62	0.45	56	14	11
B029	4.13	0.54	1.15	0.93	46	12	9
B007	4.06	0.68	1.47	0.91	44	21	4
A063	3.81	0.72	0.58	0.24	30	18	5
A022	3.76	0.49	1.30	1.14	57	11	14
B009	3.70	0.57	0.88	0.74	49	10	10
B018	3.52	0.56	0.95	1.08	51	13	12
B001	3.46	0.68	0.34	0.16	45	20	5
A017	3.45	0.52	1.02	0.83	58	10	15
B025	3.37	0.61	0.97	0.82	47	13	12
B004	3.21	0.56	1.05	0.70	43	9	16
A003	3.02	0.54	1.45	1.52	49	10	14
B038	2.72	0.59	0.85	0.77	60	6	19
A016	2.33	0.68	0.88	0.52	47	7	12
A074	2.33	0.65	0.95	0.51	45	19	6
A012	2.23	0.57	0.69	0.54	36	7	17
A046	2.22	0.61	1.27	0.76	46	8	16
A055	2.21	1.16	0.36	0.07	27	17	3
B022	1.96	0.75	0.65	0.26	31	15	5
A019	1.90	0.59	1.37	1.23	41	8	17
B040	1.76	0.61	0.68	0.36	39	8	17
B031	1.75	0.62	0.71	0.38	34	13	12
A058	1.61	0.57	0.79	0.44	34	8	17
A009	1.54	0.63	0.67	0.33	31	19	6
B020	1.05	0.68	0.71	0.34	32	4	21
A052	1.05	0.71	1.19	0.87	39	8	12
B012	0.87	0.73	1.10	0.63	36	10	15
B006	0.79	0.74	1.66	1.53	30	11	14

B002	0.76	0.69	0.62	0.32	29	10	10
B017	0.55	1.20	0.18	0.05	9	19	1
B023	0.32	0.90	0.63	0.20	27	16	4
B008	-0.43	0.91	0.29	0.10	24	16	4
A076	-0.58	0.84	0.40	0.20	24	7	8
B014	-0.60	0.79	0.48	0.16	28	5	20
A036	-0.89	0.80	0.47	0.15	32	3	22
B019	-1.42	0.72	0.66	0.30	15	14	5
A049	-1.45	0.69	0.95	0.64	29	9	15
A015	-1.53	0.90	0.59	0.17	10	15	5
A045	-1.54	0.79	0.59	0.22	19	8	12
B037	-2.21	0.72	0.34	0.15	22	12	13
B026	-2.31	0.74	1.48	1.80	19	9	10
A032	-2.44	1.17	0.18	0.05	13	17	2
A037	-2.52	1.11	0.16	0.04	28	6	19
B003	-3.15	1.00	0.22	0.08	16	4	15
A008	-3.64	0.97	0.27	0.09	16	8	12
A062	-4.26	0.94	1.34	0.52	14	4	16
B039	-4.32	1.10	0.12	0.05	7	14	6
B032	-4.40	0.74	0.58	0.24	14	7	13
B005	-4.57	0.84	0.52	0.18	8	13	7
B034	-4.58	0.92	1.08	0.27	17	5	15
A057	-4.91	0.82	0.75	0.20	17	3	22
A005	-4.92	0.81	0.43	0.16	22	7	13
B035	-5.68	0.85	0.23	0.11	13	8	12
A025	-7.01	0.84	0.49	0.18	7	8	12
A038	-7.31	0.98	0.21	0.08	8	5	15
A065	-7.54	0.83	1.03	0.34	4	4	16
B036	-7.77	0.86	0.65	0.21	10	6	12
B027	-8.17	0.95	0.38	0.11	6	6	13
B033	-8.32	0.89	0.90	0.26	4	2	18
A029	-8.72	1.07	0.67	0.18	5	1	14
A033	-8.73	1.41	0.32	0.06	15	2	13
A006	-10.38	1.10	0.17	0.06	6	3	16
A067	-11.44	1.43	0.20	0.03	9	1	24
B021	-13.15	1.93	0.11	0.02	5	0	20

English language 4 P2 – RO

Id	Measure	true.score.SE	infit	outfit	Mark	Chosen	NotChosen
A005	8.99	1.85	0.03	0.01	54	25	0
A059	6.48	0.66	1.03	0.78	57	20	3
A060	6.36	0.54	0.94	0.74	56	20	5
B021	6.23	0.54	0.96	0.65	58	20	5
B014	6.06	0.52	0.77	0.58	55	19	6
A023	5.98	0.71	0.78	0.37	53	19	5

A006	5.82	0.69	0.45	0.24	47	17	3
B022	5.49	0.52	1.29	1.45	53	17	8
B039	5.40	0.50	1.47	1.96	56	17	8
B032	5.19	2.09	0.31	0.02	24	20	0
B035	5.06	0.49	0.77	0.66	57	14	11
B036	4.92	0.64	0.38	0.19	51	18	7
B034	4.83	0.55	1.44	1.10	54	17	8
B015	4.57	0.54	1.02	0.75	52	15	9
A035	4.54	0.48	0.65	0.59	58	10	15
A038	4.40	0.57	0.82	0.58	46	14	6
A029	4.38	0.53	1.31	1.33	55	9	11
B011	4.36	0.53	1.04	0.72	48	13	12
A036	4.26	0.56	0.93	0.79	45	13	12
A061	4.17	0.87	1.22	0.74	39	11	4
A007	4.02	0.50	0.57	0.48	48	14	11
A028	3.99	0.86	0.88	0.26	44	23	2
A015	3.65	0.50	1.09	0.99	51	12	12
A033	3.61	0.54	0.56	0.35	42	16	9
A024	3.58	0.56	0.63	0.38	41	13	12
B001	3.34	0.51	1.34	1.31	32	11	14
B018	3.25	0.54	1.06	0.80	42	9	16
B009	3.12	0.58	0.64	0.42	47	8	12
B017	3.09	0.56	1.30	0.91	46	12	13
B008	2.92	0.55	0.71	0.51	45	6	14
A018	2.85	0.56	1.05	1.12	36	9	16
B007	2.75	0.58	0.66	0.41	39	11	13
B019	2.60	0.60	0.93	0.63	41	4	16
A067	2.40	0.60	1.86	1.88	52	6	18
B037	2.38	0.62	0.34	0.19	44	18	7
A064	2.18	0.67	1.11	0.99	34	8	17
B038	1.88	0.77	0.48	0.19	31	15	5
A050	1.82	0.68	0.97	0.81	29	18	7
A016	1.27	0.68	0.99	0.58	43	7	18
A063	1.16	0.96	0.36	0.10	27	16	4
B030	1.03	0.64	0.55	0.26	34	11	13
B004	0.77	0.62	0.79	0.48	43	12	12
B005	0.71	0.60	0.76	0.40	36	12	12
A062	0.64	0.77	1.20	0.53	28	10	15
A055	0.51	0.69	0.36	0.16	31	17	8
A011	0.38	0.68	0.72	0.30	30	14	11
A030	-0.10	0.90	0.22	0.09	22	16	4
B024	-0.38	0.71	1.29	0.93	30	9	14
B013	-0.76	0.71	1.60	1.32	29	6	14
B010	-0.77	0.86	0.21	0.08	28	5	20
B025	-1.22	1.17	0.14	0.05	27	13	7
A066	-1.40	0.86	1.49	0.70	32	3	22

B012	-1.64	0.76	0.30	0.15	19	11	9
B029	-1.65	0.84	0.43	0.15	15	14	6
A037	-2.20	0.87	0.77	0.32	19	8	12
B040	-2.81	1.29	0.17	0.04	13	13	7
A068	-2.88	0.81	0.44	0.18	15	9	11
A071	-2.94	0.79	0.61	0.26	24	5	15
A022	-3.31	0.67	0.78	0.38	17	7	18
A070	-3.76	1.07	0.23	0.07	16	10	10
B027	-3.84	0.83	1.51	1.02	16	3	15
A003	-3.90	0.88	0.49	0.15	13	17	3
B003	-3.92	0.90	0.47	0.15	10	15	5
B033	-4.36	1.16	0.42	0.08	17	5	15
A073	-4.69	1.09	0.32	0.08	14	4	16
B020	-4.77	0.73	1.03	0.44	22	6	19
B023	-5.62	0.84	0.69	0.25	6	11	9
A004	-6.02	0.83	0.54	0.20	5	10	10
B028	-6.12	0.99	0.38	0.11	14	4	16
A032	-6.69	1.00	0.95	0.34	4	7	13
B006	-7.38	0.86	0.26	0.11	9	7	13
B026	-7.61	0.76	0.52	0.25	7	7	13
A020	-8.38	0.81	0.97	0.46	10	4	16
A001	-8.47	0.70	0.99	0.51	9	6	14
A002	-9.03	0.67	0.97	0.49	7	5	15
A026	-9.45	0.79	0.58	0.21	8	3	17
B002	-9.52	0.66	0.76	0.38	8	5	15
B016	-9.69	0.69	1.19	0.62	5	4	16
A009	-10.72	1.00	0.73	0.20	6	1	19
B031	-11.35	1.92	0.10	0.02	4	0	20

English language 4 P1 – PCJ

Id	Measure	true.score.SE	infit	outfit	Mark	Chosen	NotChosen
A044	5.89	0.98	0.72	0.17	62	25	1
A041	5.40	0.82	0.89	0.23	53	24	2
A050	5.26	0.83	0.83	0.22	59	24	2
B015	4.98	0.71	0.78	0.28	58	22	3
B029	4.97	0.75	0.75	0.23	46	22	3
A068	4.89	0.83	0.91	0.23	43	24	2
A030	4.76	0.75	0.92	0.45	51	23	3
A048	4.75	0.67	0.67	0.25	60	22	4
B028	4.72	0.8	1.25	0.37	59	22	3
B011	4.64	0.71	0.75	0.28	56	23	3
A060	4.48	0.73	0.66	0.21	56	22	4
B038	4.45	0.77	1.42	0.61	60	22	3
B024	4.26	0.64	1.07	0.8	53	20	5
B010	4.18	0.88	1.42	0.52	57	22	3

B018	4.09	0.73	0.93	0.39	51	22	4
A073	3.59	0.77	1.11	0.91	61	21	4
B013	3.58	0.72	0.43	0.17	62	23	4
A017	3.02	0.73	0.55	0.19	58	23	4
A003	2.87	0.62	0.68	0.3	49	20	8
B016	2.71	0.65	1.16	0.66	41	20	6
B025	2.47	0.71	0.45	0.17	47	18	7
A016	2.42	0.67	1.29	0.69	47	18	7
A019	2.39	0.61	0.81	0.39	41	17	9
B001	2.33	0.71	0.81	0.5	45	15	10
A046	2.32	0.63	0.65	0.29	46	18	8
B007	2.16	0.65	0.57	0.25	44	16	9
B040	2.13	0.69	0.41	0.17	39	16	10
B030	2.12	0.77	0.97	0.27	61	19	8
A022	2.08	0.58	0.89	0.64	57	17	9
A052	2.03	0.66	1.01	0.42	39	22	7
A074	1.84	0.57	0.95	0.79	45	13	13
B009	1.75	0.68	0.76	0.29	49	16	10
B004	1.7	0.64	0.83	0.37	43	17	9
B006	1.64	0.61	1.32	0.81	30	16	10
A056	1.5	0.66	0.37	0.17	44	14	13
B012	0.82	0.68	1.3	0.9	36	17	8
A009	0.73	0.68	0.92	0.9	31	12	14
A058	0.56	0.68	0.52	0.23	34	12	14
B023	0.54	0.77	0.96	0.36	27	15	11
B014	0.36	0.69	0.79	0.36	28	14	11
A012	-0.09	0.79	0.51	0.16	36	12	14
B031	-0.19	0.68	0.59	0.24	34	13	12
B020	-0.24	0.68	0.59	0.24	32	12	14
B017	-0.29	0.68	1.74	0.87	9	10	16
A063	-0.36	0.65	0.42	0.2	30	9	16
B037	-0.41	0.7	0.44	0.17	22	13	13
B008	-0.48	0.64	0.72	0.32	24	9	16
B022	-0.52	0.72	0.59	0.21	31	11	14
B002	-0.56	0.79	0.48	0.16	29	13	12
A045	-1.33	0.69	0.63	0.26	19	7	19
A036	-1.36	0.77	0.93	0.5	32	10	16
A055	-1.56	0.92	0.16	0.06	27	11	15
A037	-1.58	0.68	0.72	0.58	28	12	14
A049	-1.61	0.7	0.81	0.35	29	7	18
B026	-1.83	0.76	0.84	0.63	19	11	14
A076	-2.09	0.72	0.66	0.25	24	9	16
A005	-2.62	0.67	0.45	0.18	22	9	19
B034	-2.79	0.66	0.93	0.47	17	6	18
B032	-2.96	0.84	1.07	0.72	14	5	20
B019	-3.07	0.88	0.27	0.08	15	4	21

A062	-3.27	0.65	0.96	0.6	14	10	15
B003	-3.32	0.73	0.44	0.16	16	6	20
A015	-3.59	0.72	0.81	0.5	10	5	21
B036	-3.69	0.8	0.76	0.22	10	4	21
B005	-4.01	0.69	0.63	0.24	8	5	21
A008	-4.02	0.81	0.4	0.12	16	5	21
A032	-4.12	0.85	0.31	0.09	13	5	21
B039	-4.2	0.8	1.08	0.33	7	3	25
A057	-4.27	0.76	0.44	0.15	17	4	22
B035	-4.36	0.67	0.93	0.71	13	4	22
A067	-4.77	0.77	0.9	0.28	9	4	21
A025	-5.07	0.79	0.97	0.29	7	3	23
A006	-5.34	0.85	0.5	0.14	6	2	24
A033	-5.37	0.75	0.92	0.47	15	3	23
A038	-5.46	1.03	0.99	0.17	8	2	25
A029	-5.67	0.9	1.47	0.74	5	2	24
B021	-5.8	1.06	0.82	0.14	5	1	24
A065	-6.31	1.01	0.63	0.12	4	1	25
B027	-6.33	1.07	0.39	0.07	6	1	25
B033	-6.45	1.05	0.9	0.17	4	1	24

English language 4 P2 – PCJ

Id	Measure	true.score.SE	infit	outfit	Mark	Chosen	NotChosen
B035	7.73	1.93	0.12	0.01	57	24	0
B021	7.27	1.90	0.08	0.01	58	24	0
A005	6.16	1.35	0.13	0.03	54	25	1
A060	5.26	0.90	0.64	0.19	56	23	2
B022	4.69	0.79	0.46	0.15	53	23	2
A059	4.66	0.82	1.30	0.90	57	23	2
A038	4.60	0.85	0.47	0.12	46	24	3
B036	4.54	0.79	1.51	0.93	51	23	3
B014	4.49	0.74	0.64	0.24	55	21	4
A015	4.17	0.88	1.16	0.40	51	23	2
A007	4.12	0.77	1.17	0.83	48	22	4
A035	3.66	0.70	0.45	0.17	58	22	4
A023	3.63	0.61	0.80	0.40	53	20	6
B039	3.49	0.65	0.78	0.53	56	22	5
B034	3.48	0.65	0.46	0.21	54	20	5
A006	3.48	0.72	0.81	0.30	47	20	6
B008	3.07	0.66	0.70	0.29	45	22	5
B015	2.98	0.60	1.04	0.76	52	16	9
B004	2.54	0.68	1.28	0.66	43	18	8
A061	2.53	0.70	0.50	0.19	39	21	5
A067	2.47	0.72	0.77	0.29	52	19	8
A036	2.33	0.69	1.15	0.59	45	17	9

A064	2.22	0.67	0.71	0.36	34	19	7
B037	2.11	0.57	0.67	0.37	44	17	10
B019	2.11	0.67	0.59	0.24	41	18	7
A033	2.03	0.63	0.59	0.27	42	17	9
B011	2.01	0.72	0.27	0.13	48	18	7
B017	1.96	0.62	1.38	1.10	46	16	9
A018	1.88	0.61	0.90	0.40	36	19	8
A024	1.83	0.61	0.93	0.54	41	19	7
A028	1.78	0.58	1.04	1.02	44	16	11
A029	1.62	0.64	0.97	0.56	55	15	11
B018	1.53	0.54	0.91	0.97	42	14	12
B009	1.40	0.65	0.60	0.27	47	17	7
B038	1.36	0.64	1.67	1.36	31	15	10
B001	1.26	0.61	0.88	0.63	32	18	8
B032	1.23	0.71	0.68	0.24	24	16	9
A055	0.88	0.62	0.71	0.39	31	12	14
A016	0.88	0.64	0.77	0.35	43	14	12
B025	0.78	0.66	1.34	0.91	27	15	10
B030	0.77	0.60	0.70	0.37	34	15	10
B024	0.62	0.72	1.04	0.41	30	16	11
A062	0.48	0.69	0.54	0.22	28	14	11
A063	0.36	0.64	0.68	0.38	27	11	15
A050	0.32	0.60	0.75	0.41	29	10	16
B007	0.03	0.60	0.72	0.52	39	9	15
B005	-0.02	0.63	0.51	0.24	36	13	14
A066	-0.26	0.74	0.94	0.50	32	15	9
A030	-0.52	0.67	0.92	0.53	22	11	15
B010	-0.56	0.71	0.56	0.24	28	11	13
A011	-0.82	0.75	0.35	0.13	30	7	20
B020	-0.91	0.69	0.67	0.34	22	10	17
A071	-1.19	0.83	0.92	0.69	24	9	17
B013	-1.38	0.78	0.49	0.20	29	8	17
B012	-1.84	0.99	0.11	0.05	19	8	17
A073	-1.92	0.94	0.14	0.06	14	7	19
A022	-2.51	1.01	0.18	0.05	17	6	20
B040	-2.57	0.87	0.44	0.12	13	7	19
B027	-3.01	0.94	0.79	0.17	16	4	21
B033	-3.19	0.79	0.60	0.18	17	6	18
A068	-3.22	1.03	0.41	0.08	15	5	21
A070	-3.23	0.72	0.47	0.17	16	8	18
A037	-3.26	0.76	1.00	0.44	19	5	21
B029	-3.41	0.81	0.78	0.28	15	8	18
B006	-3.68	0.78	1.00	0.49	9	7	18
A003	-4.20	0.94	0.36	0.08	13	3	24
A004	-4.37	0.83	0.39	0.11	5	3	24
A020	-4.43	1.13	0.34	0.06	10	2	24

B003	-4.62	0.93	0.31	0.08	10	5	20
B023	-4.70	0.98	0.35	0.08	6	3	22
A009	-4.88	0.86	0.98	0.29	6	2	24
A002	-4.93	0.97	0.38	0.08	7	2	24
B028	-5.20	0.84	1.20	0.45	14	4	22
B016	-5.25	0.94	0.43	0.10	5	2	23
A032	-5.74	0.95	1.05	0.36	4	3	24
A001	-6.08	0.91	0.60	0.14	9	2	26
B026	-6.59	1.32	0.17	0.03	7	1	25
B002	-7.11	1.88	0.07	0.01	8	0	27
A026	-8.00	1.27	0.19	0.03	8	1	25
B031	-9.22	2.01	0.22	0.01	4	0	25

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

0300 303 3344
public.enquiries@ofqual.gov.uk
www.gov.uk/ofqual

December 2019 **Ofqual/19/6575**