Pearson English International Certificate:

Establishing Comparable Standards Across Test Modes.

August 2022

Dr. Rose Clesham

Lauren Miller

Executive Summary

This paper describes the processes and procedures that were used to establish the standards of the new Pearson English International Certificate (PEIC) computer-based test. There is an existing PEIC paper-based test and test takers will be able to choose between the computer and paper-based modes. It is important therefore that the standards that are set and operationalised on both modes are comparable. Both test modes test the same construct, English Language proficiency as described by the underlying language framework of the Common European Framework of Reference for Languages (CEFR). Both the paper and computer-based tests have been designed to test the four communicative skills of reading, writing, speaking and listening through the key indicators described at each level of the CEFR.

Comparable standards across the two modes have been established using a number of strands of evidence, leading to a formal standard setting exercise. The strands of evidence included:

- The use of comparative judgement (CJ)
- The creation of a cognate archive with agreed threshold performance descriptors
- A modified Angoff exercise
- A Beta testing phase including common test takers
- Sense checking activities
- A standard setting exercise.

These activities will be described in this paper. To further aid the comparability work and evidence, the expert rating panel was populated by the senior members of the paper-based examining team.

The standards that have now been set and operationalised will be regularly monitored as part of ongoing validation exercises.

Background

The Pearson English International Certificate (PEIC) paper-based test, formerly known as the Pearson Test of English General (PTE General) has been available for many years and is offered across the six levels of language proficiency set out in the Common European Framework Reference for Languages (CEFR), A1 to C2, (sometimes referred to the levels of basic to proficient user (Council of Europe, 2020)).

The paper-based test is offered at present in seven series a year, with a unique test available at each level in each series. These tests are constructed manually for each series drawing from an item bank where items have been written targeted at specific CEFR levels. The tests are marked by teams of examiners for each level. The standard for each CEFR test level was originally established by expert judgement and awarding takes place after every series to maintain that standard using classical test statistics and qualitative judgements by a panel of senior examiners. The judgemental thresholds determined each time are at pass and distinction, and the merit threshold is determined arithmetically at the midpoint between pass and distinction. A fallback grade is also now set arithmetically for levels A2-C2, providing a pass grade at the level below the test taken.

The paper-based test does not use common or calibrated items in test construction and so tests are not psychometrically equated. Therefore, the standard for each test cannot be maintained using item response theory (IRT) and so judgemental methods are used instead.

To broaden the provision of the PEIC test, a computer-based version at each CEFR level has been developed over the last couple of years. These computer-based versions can be administered in a test centre, and now also at home using human and AI enhanced remote proctoring. The computer-based tests are underpinned by psychometric measurement. Test items will be banked with known difficulty values, and standards set using IRT ability estimates.

Introduction

This paper describes the form of the new computer based PEIC test and how the standards were set in order to align with those of the established paper-based tests. The scope of this report is limited to the standard setting alone. It will not report on test development or any other aspects of the PEIC computer-based test life cycle.

The computer-based test was developed with a number of test forms targeting each CEFR level. Using psychometric measures, common items link the tests within and across levels, which allows all tests to be placed on the same psychometric scale.

The data collected during the beta testing phase was used to calibrate this psychometric scale and lay the groundwork to transition to a fully item-banked system. Now that the computer-based tests are live, new items will be seeded within test forms and the calibrated items stored in an item bank. Once the item bank is sufficiently large, fixed test forms will be replaced by a linear on the fly test (LOFT) model, whereby unique tests are constructed using

the known item difficulty values from the seeding process, thus ensuring the unique test forms are identical in difficulty.

The test is not adaptive and test takers are entered at and sit a specific CEFR level test. Each test taker's result is an IRT ability estimate which is then converted to a test score result. As the computer-based tests have been determined psychometrically using a calibrated item bank, testing can be continuous, thereby dispensing with examination series, and grades can be awarded without the need for ongoing awarding meetings.

When fully operational, engine training will take place using human raters' expert judgement, thereby, making the automated marking of the computer-based test possible. Further information about AI scoring can be found on the Pearson Test of English website.

This paper describes the processes and procedures that were worked through in order to translate the standards established in the paper-based tests across onto an IRT logit scale used on the computer-based tests.

The Construct

As described above, the PEIC tests have been established to provide flexibility for test takers. The paper and computer-based test modes have differing test designs, based on the different affordances of paper and computer-based items. In order to offer differing modes for the same qualification, the construct tested must be the same across test modes. In the context of PEIC, this has been made possible by the underlying language framework of the CEFR. Both the paper and computer-based tests have been designed to test the four communicative skills of reading, writing, speaking and listening through the key indicators described at each level of the CEFR. This has been established by extensive training and standardisation of item writing and review processes.

Test Mode Comparability Processes and Procedures

This section will describe the processes, procedures and methodologies implemented in order to establish comparability of standards across test modes.

Figure 1 below shows the equating measures used and the general timeline for these activities:

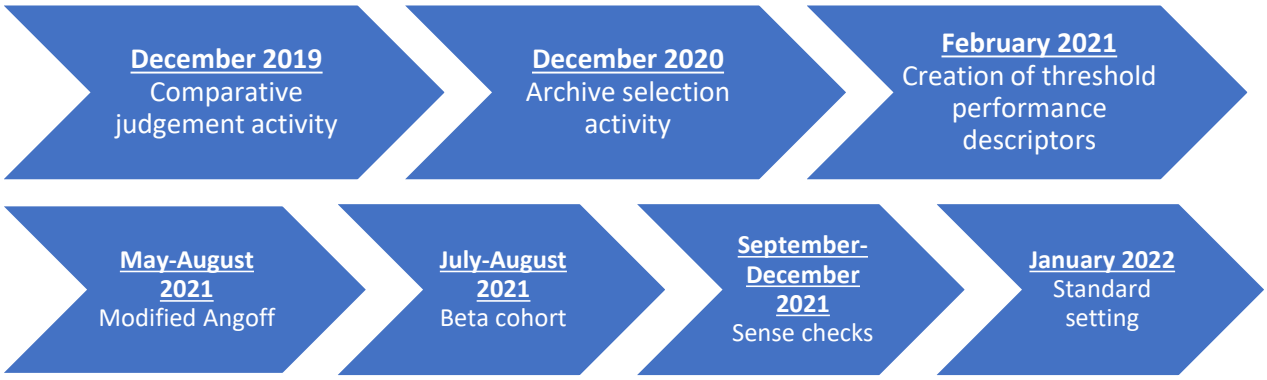

Figure 1: Timeline of equating measure activities

Each of these steps will be described:

- A comparative judgement (CJ) activity to understand the perceived item difficulty
- A cognate archive selection activity, which facilitated the creation of threshold performance descriptors
- A modified Angoff activity
- Common test takers sitting both test modes (Beta 1)
- Sense checks
- The standard setting itself.

The paper-based test senior examining team consists of one Chair of Examiners, three Chief Examiners and seven Principal Examiners. As this team have been in place for many years, their insights were invaluable in translating the established standards of the paper-based tests to the new computer test forms.

Comparative Judgement Activity

When setting out to understand the potential differences in the test modes, the first priority was to understand the comparability of test item difficulty. The new PEIC items were ready before the test forms were created, so a comparative judgement (CJ) activity was designed to investigate the level of difficulty of items across the paper and computer-based tests.

Several methods exist to quantify the perceived difficulty of items and test forms, and the majority of these are based on absolute judgements. In absolute judgement methodologies, judges are asked to review an item and assign a score that describes the item's difficulty on a fixed scale (e.g. the CEFR). Comparative judgement (CJ) offers an alternative perspective to these traditional methods (Thurstone, 1927). In CJ tasks, judges are presented with a pair of items and asked to identify the more difficult item. Because the judgements are quick and intuitive, a large number of comparisons can be generated. The judgements are then analysed statistically using the Bradley-Terry model (Bradley & Terry, 1952) and Rasch (Rasch, 1960/1980), to establish a scale of difficulty/ability. CJ tasks can efficiently produce a reliable scale because the scale is based on an aggregation of large numbers of simple judgements from a wide range of judges.

In this CJ study, we compared the difficulty of a large number of paper and computer-based items. Judges were given this simple set of instructions to guide their judgements:

'Two test items will appear on your screen and you will decide which one is the more difficult item.'

23 subject experts made comparisons of items from both test modes, across the full CEFR ability range and the four primary English skills (listening, reading, writing, speaking) assessed in the PEIC tests to establish a rank order of the perceived difficulty of the computer-based test and paper-based test items. This activity did not require complete test forms, so could take place relatively early in the development life cycle.

The study was broken into five activities, one for each of the skills as well as an activity comparing all items from all skills. All five activities were conducted in Pearson's own assessment research comparative judgement platform, as this platform includes the functionality required to play audio clips for listening items. The platform employs a pairing algorithm so that pairs of items were selected at random, with preference given to items and pairs of items that had not yet been judged.

A total of 65,854 judgements were made in this CJ, with the following breakdown per skill:

- 15,485 judgements for listening
- 15,166 judgements for reading
- 784 judgements for writing
- 3,138 judgements for speaking
- 31,281 judgements for the combined skills activity

Each item was judged between 65-92 times across the five activities. The approximate average time taken for all judgements was 21 hours.

All five activities produced highly reliable scales, with reliability greater than 0.95 for each activity, meaning there was extremely high consensus amongst judges to build a reliable rank order and scale of difficulty estimates.

The output showing the comparison of perceived difficulty across all items in the combined skills activity is provided in Figure 2 below:

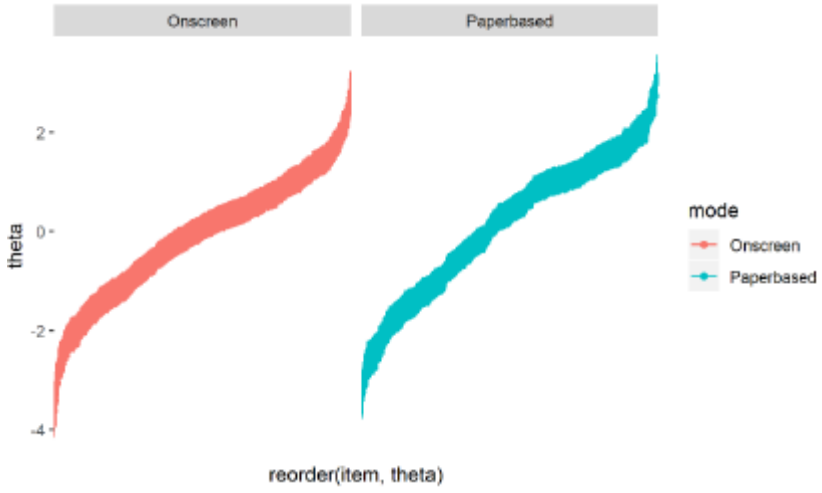

Figure 2: CJ output showing comparison of perceived item difficulty

Figure 2 shows that the items for two modes of assessment were judged to cover the same range of ability in a broadly similar pattern. Further analysis of the four skill areas showed similar results.

Observations from this study indicated that:

- The study yielded highly reliable scales of difficulty and a high level of judge consensus.
- The item difficulty estimates are broadly comparable across the modes of testing.

The comparative judgement activity provided a useful first look at the items for the computer-based test and was the first step is understanding how the two test modes compared.

Cognate Archive Selection

The use of an archive of scripts on the grade boundaries from a previous series is typical in UK awarding (Taylor & Opposs, 2018; Davenport, 2008). For the first award of a new qualification, a cognate archive from a similar qualification is used. The purpose of an archive is to provide a qualitative indication to the judges of the quality of performance expected at that grade threshold.

In December 2020 the paper-based test senior examining team were provided with speaking and writing item responses from test takers who had sat the paper-based test and were on the pass and distinction grade thresholds. From those responses the senior examiners were asked to select those that best demonstrated threshold performance for each grade and then produce brief commentaries to highlight the characteristics of threshold performance in each response.

At each stage the Chair of Examiners was consulted and given the opportunity to veto responses. Once the final commentaries were drafted, the Chair consolidated and standardised the commentaries and used them to create comprehensive threshold performance descriptors to be used in subsequent expert judgement led activities.

Cognate archives and performance descriptors are both well-established methods of articulating key indicators for standard setting purposes. They can be added to throughout the lifecycle of a qualification and ensure that in standard setting judges adhere to a common understanding of performance standards.

Modified Angoff

Due to the Covid-19 pandemic, trialling and calibrating the computer-based items was not possible in advance of the live beta implementation of the test. In such circumstances, Angoff activities provide invaluable evidence to inform the location of mark ranges when standard setting activities take place. Although the computer-based test thresholds would subsequently be placed on a psychometric scale instead of a proportion of marks out of the total, raw marks did exist in the initial test forms to facilitate human scoring for training AI. Therefore, notional grade thresholds for each level and judgemental threshold (pass and distinction) in raw marks were identified before the standard setting took place and were used as initial points of triangulation for the pre-standard setting analysis.

Following construction of a number of fixed computer-based test forms, a modified Angoff activity took place to help to identify notional grade thresholds. As there are common items throughout the test forms as an internal equating measure to place all computer-based tests on the same psychometric scale, the judges rated pools of item types instead of complete test forms, to avoid making the same judgements on the same item more than once.

The judges were split into groups and asked to rate items in their allocation. This allocation included items specific to each judge group as well as common items across all judges. Judges were asked to make judgements for the pass and distinction thresholds at each level for each item, making 12 judgements per item in total. When rating, the judges were asked to estimate the probability of a borderline test taker at each threshold at each level getting the item correct. To assist in making this judgement, the judges were instructed to ask themselves one of two questions, depending on the item type:

Q1. If there were 100 similar items, how many would a borderline test taker score correctly at each threshold, at each level?

An alternative phrase used to explain the instruction was:

Q2. If each item were worth 100 marks, how many marks would a borderline test taker score at each threshold, at each level?

The first question (Q1) applied to multiple choice questions and other closed items. The second question (Q2) applied to open items, for instance essay questions. Both questions yielded the same outcome: the perceived probability of a borderline test taker at any given level getting the item correct. From this, the ratings were converted into mean scores for each item and overlayed onto the test forms. This would indicate potential notional boundary thresholds based on the raw marks of each test and be the starting point for standard setting.

Following data cleaning, extreme outlier judgements that would skew average ratings were removed. The criteria for this were any judgements more than 20 points away from the average rating. After cleaning, the number of ratings used for the analysis was 81,679.

Further analysis into the standard deviations (SD) for all ratings at every threshold on each item was carried out. All standard deviation values over 20 were flagged for the purpose of secondary analysis when identifying the notional thresholds. Larger SDs indicate less judge consensus about the difficulty of the item. These items were not removed from the overall notional threshold analysis, with all items retained. However, the SD information was used to anticipate low consensus, provide helpful context, and facilitate discussion during the standard setting activity.

This judgemental exercise was subsequently used alongside empirical performance data and expert inspection of responses to set threshold scores.

Beta 1 Cohort (common test takers)

In an attempt to establish a statistical equating measure, a group of test takers were recruited to sit the paper-based and computer-based tests. This cohort was named Beta 1. As concordance and linking studies have shown, one key control variable for this type of study is to keep the time between taking tests as short as possible. Beta 1 sat the paper-based test in either March, May or June 2021 and had been awarded a result. The Beta 1 cohort then sat the computer-based version throughout July and August 2021. This cohort were not awarded grades from the computer-based tests, and they knew that their test taking

experiences were providing modal validation evidence. They were each offered a gift voucher as an incentive to make a genuine attempt at sitting the computer-based test.

In some cases, up to five months had elapsed between test takers sitting the paper-based test and sitting the computer-based test, so while the exercise provided a range of useful feedback in terms of test taking performance and attitudes, the correlation evidence between the two tests was limited.

The numbers of test takers at each level who had sat both test modes are provided below:

Level	Number of test takers
A1	9
A2	20
B1	26
B2	54
C1	5
C2	-

Table 1: Beta 1 cohort test taker numbers

There were no C2 level test takers who sat both the paper-based test and computer-based test and the number of A1 and C1 test takers were very low.

While for some levels the correlation coefficient values indicated a good correlation between modal scores, some levels had low values for this measure. A correlation coefficient value of 0.7 generally indicates a good relationship between two sets of values. The correlation coefficient values for each level are provided below:

Level	Correlation coefficient value
A1	0.60
A2	0.80
B1	0.32
B2	0.57
C1	0.78
C2	-

Table 2: Correlation coefficient values for Beta 1 cohort test scores

Multiple factors could account for the poor correlation between test mode scores. As mentioned above, a key factor was the time elapsed between taking the two modes of test with further learning opportunities. In addition, unfamiliarity with the computer-based items and format was also a contributory factor. As this research phase was in beta mode itself, there were no preparation of familiarisation materials available for test takers, and the test takers knew that they would not be awarded a result, which may have affected their motivation.

Although this attempt to establish a statistical equating measure was not successful, the Beta 1 phase was instrumental in being able to test the notional thresholds from the modified Angoff at an early stage through sense checks, which are outlined in the following section.

After this Beta 1 phase, the computer-based tests were made available to a Beta 2 cohort in a soft launch of the computer-based test. These test takers were sitting tests under normal live test conditions. However, at this Beta 2 stage, their results could not be awarded until enough data was collected and a standard setting activity undertaken.

Sense Checks

In non-psychometric paper-based awarding procedures, sense checks are routinely used following statistical analysis to assist in identifying an appropriate script inspection range to determine where to set threshold scores. For the computer-based test, two sense checks took place, the first of which took place immediately following the Beta 1 cohort window, and the second immediately prior to the standard setting. The Chair of Examiners for the paper-based test completed both sense checks.

For the first sense check the Chair of Examiners was instructed to review responses from the Beta 1 test takers who were on or near the notional thresholds from the modified Angoff activity, as well as test takers who had been on or near the awarded grade boundaries from the paper-based test. The Chair was asked to determine whether the test taker was below threshold, threshold or secure, for both pass and distinction. There were no A1 test takers in the pass range, therefore these criteria could not be tested for that threshold.

The purpose of the second check was to ensure the ranges of test taker work that would be judged in the standard setting activity were appropriate. Further information regarding this is in the standard setting section below.

Standard Setting

In December 2021, a psychometric recalibration was carried out to provide up to date empirical data of the ability estimates of all computer-based test takers since live implementation. This was based on a significantly larger number of test takers compared to the Beta 1 cohort. The resulting scale was used in the standard setting exercise, and cut scores were determined from this scale, thus enabling reliable threshold scores across test forms.

The outputs from the first sense check and notional thresholds from the modified Angoff activity were used to identify preliminary ability estimates for the inspection ranges. Following the recalibration, the Chair of Examiners carried out a second sense check to confirm the inspection ranges were appropriate to carry forward. The inspection ranges were based on ability estimates, not raw marks. Within each inspection range test takers with even ability profiles across all four skills were prioritised for inspection.

The item types included in the inspection ranges were items that assessed the productive skills of speaking and writing. Items that solely assessed listening and reading were not included as there was little test taker 'performance' in these items to make judgements on.

Table 3 below shows the item types that were included in the inspection.

Description of item	Skills assessed
Read a passage and reproduce it	Reading and writing
Read a passage aloud	Speaking and reading
Listen to a sentence then repeat it	Listening and speaking
Write an essay on a given topic	Writing
Look at an image and describe it	Speaking

Table 3: PEIC computer-based test items included in inspection ranges

The standard setting activities took place in January 2022. All judges participated in a pre-standard setting familiarisation meeting to ensure a common understanding of the objectives and processes of this activity

The senior examining team from the paper-based test were invited to participate in the computer-based standard setting, although the membership for each level changed depending on senior examiner availability. At a minimum the Chair of Examiners, Chief Examiner and Principal Examiners for each level on the paper-based test were included in the judging panel for the same level on the computer-based test.

The standard setting methodology used is commonly referred to in the UK as 'awarding', or the body of work method (Opposs & Gorgen, 2018). Judges completed a first round of inspection independently on two sets of test takers (pass and distinction) for each level. The judges rated the work of a range of test takers as below threshold, threshold or secure. Following consolidation of these judgements a grey area was defined. The grey area is the area of performance where threshold performance is seen and within which the cut score can be set. In UK awarding procedures, this grey area would range from a score where some threshold performance was seen in some test takers to a score where all test takers showed threshold performance.

A larger grey area than we would see in a traditional paper test was anticipated in the computer-based test as it was a new test, and some uncertainty was expected. A discussion of all the test takers in the grey area took place before a second round of judging, focusing on test takers in the grey area only. The grey area was refined and the final outputs from the standard setting were ranges in which the cut scores for pass and distinction would be placed.

Once the process outlined above had taken place for all levels, the psychometric ranges for the grey areas were established. This step needed to take place once all levels were complete as all the cut scores would be on the same psychometric scale, and so it was necessary to ensure that as well as each cut score being in its respective grey area (and in line with pre-standard setting evidence, e.g. notional thresholds from the modified Angoff and sense check outcomes), the thresholds needed to be sequential, i.e. A1 pass, A1 distinction, A2 pass, A2 distinction and so on. The cut score placement is shown in visual form in Figure 3 below:

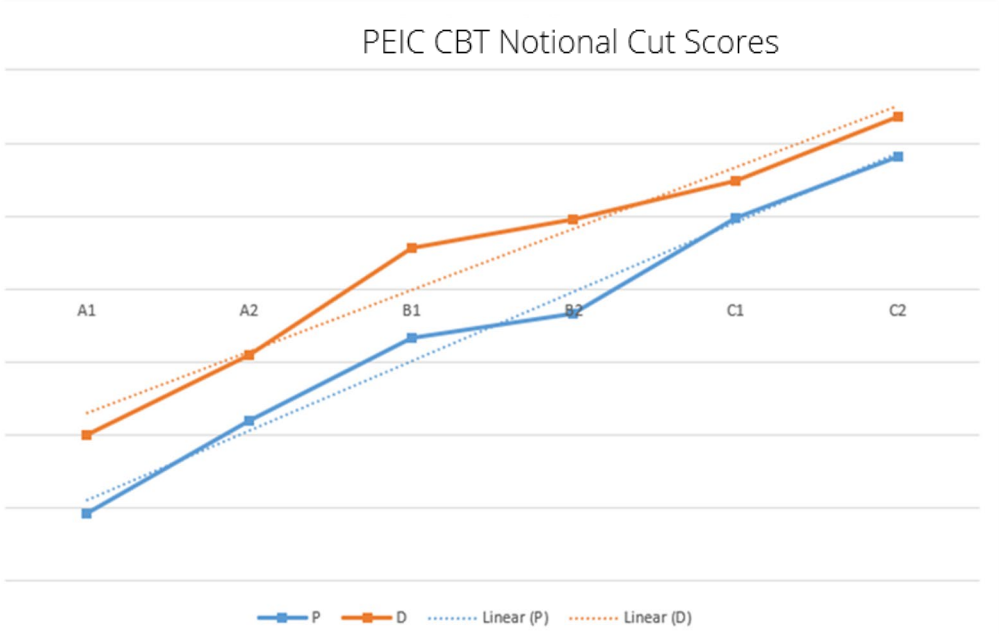

Figure 3: PEIC computer-based test cut score placement. From a Pearson internal document by J. Strangeways, 2022. Reproduced with permission.

As discussed earlier, the merit and fallback grades were set arithmetically. The fallback grade awarded for each level was the pass threshold for the level below. All levels except A1 had a fallback threshold. A total of 23 grade thresholds were set in this standard setting. The boundaries recommended following this standard setting activity are now operational. Ongoing standards verification work will take place to ensure the standard across the test modes remains comparable.

Standards Verification

Ongoing standards verification work is necessary to ensure the cut scores derived from the standard setting activities continue to represent the threshold standard, and that the threshold standard remains aligned to that of the paper-based test. This means that the computer-based test psychometric cut scores are monitored over time and may change, given empirical and expert judgemental evidence.

The standards of the computer-based test will be verified from the point where there are sufficient test takers in the refined grey areas from the standard setting to carry out standards verification work. It is likely that standards verification will occur at different times for each level, depending on the entry numbers. Monthly analysis of test taker entries and test scores is now underway to facilitate this.

The methodology for the standards verification will closely follow the procedures of the standard setting activities, using the same panel of judges.

References

Bradley, R. A., & Terry, M. E. (1952). Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika, 39*(3/4), 324–345. Accessed 27 June 2022 from : https://doi.org/10.2307/2334029

Council of Europe. (2020). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Accessed 18 April 2022 from: https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Davenport, C. (2008). Evaluation of the February 2008 confirmation method awarding trial, Manchester: AQA Centre for Education Research and Policy. Accessed 15 June 2022 https://filestore.aqa.org.uk/content/research/CERP_RP_CD_01052008.pdf

Opposs, D; Gorgen, K (2018) What is Standard Setting? In J. Baird, T. Isaacs, D. Opposs and L. Gray (Eds) *Examination Standards: How Measures and Meanings Differ Around the World* (p.54-76), UCL Institute of Education Press

Opposs, D. and Taylor, R., (2018) Award in England: A Levels in Baird, J., Isaacs, T., Opposs, D and Gray, L. (Eds), *Examination Standards: How Measures and Meanings Differ Around the World* (p.100-113) UCL Institute of Education Press, University College of London

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. Nielsen & Lydiche.

Strangeways, J. (2022) *PEIC CBT Notional Cut Scores*, internal paper, Standards and Awarding, Pearson

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286. Accessed 27 June 2022 from https://doi.org/10.1037/h0070288