A comparability study using a rank-ordering methodology at syllabus level between examination boards

L.W.K. Yim and S.D. Shaw
University of Cambridge International Examinations (CIE), Cambridge Assessment
1 Hills Road, Cambridge, England, CB1 2EU, UK

Abstract

As part of the on-going programme of work to ensure the equivalence of standards of similar qualifications across different awarding bodies, both national and international, a new methodology for comparison has been trialled. A rank-ordering method [1,2] has been used relatively effectively to compare standards across boards at component level within a syllabus; however no attempts have been made to conduct the comparison at syllabus level using similar methods. The aim of this study is to conduct a modified approach of the existing rank-ordering method and to compare the examination standards of a syllabus between two boards at syllabus level. Question papers, mark schemes, syllabus specifications and candidates' scripts for all components for the same examination session were collected from both boards. These were then evaluated by external consultants and the resulting data were analysed using the multifacet Rasch modelling technique. The methodology of the study, the research outcome and feedback from examiners are described in this paper.

1. Introduction

In England, there are a number of examination boards offering public examinations which lead to the same qualifications, i.e. GCSE and GCE A Level. Although each examination syllabus must conform to general qualifications criteria approved by the examination regulator, and generally also to a common core of subject content, the syllabuses may differ between boards in other respects. A crucial question of whether it is easier to pass a particular examination with one board rather than another arises. In fact, this issue is not limited to England alone, but extends to overseas countries where candidates sit for examinations which are claimed to be equivalent qualifications to the GCSE and GCE A Level.

To ensure the equivalence of standards of similar qualifications across different awarding bodies, several research programs have been conducted, most of which compare only examination standards qualitatively or pseudo-quantitatively between examination boards. A rank-ordering method [1, 2] has been used relatively effectively to compare standards across boards quantitatively at component level within a syllabus[1]; however no attempts have been made to conduct the comparison at syllabus level using similar methods. The aim of this study was to conduct a modified approach of the existing rank-ordering method and to compare the examination standards of a syllabus between two boards at syllabus level. The rationale behind conducting research at syllabus level is that quantitative results can generally help inform grading decisions in terms of threshold adjustment of an entire syllabus where there is a need to align standards with another examination board. The materials used in this modified rank-ordering approach were question papers, mark schemes, syllabus specifications and candidates' scripts for all components for the same examination sessions from both examination boards. These were then evaluated by

[1] In CIE, a syllabus usually comprises different options, e.g. A, B and C. A combination of components will make up different options, e.g. Option A comprises components 1 and 2, Option B comprises components 1 and 3 and Option C comprises components 2 and 3 and so on. If there is only one option in the syllabus, the terms syllabus level and option level are interchangeable.

external consultants to generate rankings of the pseudo-candidates' scripts[2]. The resulting data were analysed using multifacet Rasch and the difference in standards between two boards was deduced from graphical representations. The methodology, the research outcome and consultants' feedback of this modified approach are described below.

2. Understanding comparability

Comparability in this context is concerned with the application of the same standard across different examinations [3]. Comparable examinations have to be at the same standard. However, what is meant by 'examination standard' and what kinds of comparability are expected [4]? The purpose of inter-board comparability studies is to compare standards across different awarding bodies. In making this comparison, it is important to distinguish between *content standards* and *performance standards* described by Hambleton:

> *"Content standards refer to the curriculum [or syllabus/specification] and what examinees are expected to know and to be able to do … performance standards communicate how well examinees are expected to perform in relation to the content standards"* [5].

Based on the approach of Hambleton in distinguishing between content standards and performance standards, it is important to differentiate the two standards particularly as the distinction determines the suitability (or otherwise) of a potential methodology.

3. Inter-board comparability studies

Inter-board comparability studies traditionally investigate three strands of activity.

- A review of the demands placed on candidates by the syllabus specifications, mark scheme and question paper. Essentially this is an exercise in comparing the content standards of different syllabuses. Judgements about content standards invariably rely on values and a requirement for expert participants to fully familiarise themselves with each awarding body's syllabus/specification materials.

- A cross-moderation exercise in which examiners compare pairs of scripts and decide which one demonstrates better quality. Essentially, this is an activity in comparing performance standards of the different boards. This exercise draws heavily on the expertise of senior examiners to judge the quality of examinees' work taking into account the demand placed upon them by the individual syllabuses/specifications, question papers and mark schemes. Additionally, a ratification method has been employed whereby judges are expected to identify whether each script is below, at, or above the borderline region.

- A statistical analysis such as multilevel modelling can be carried out by including flexible analysis techniques allied with nationally available value-added data, together with a wide variety of background information at every level, for example, *candidate level*: candidates' personal data, prior attainment, current grades; *school level*: types of school (e.g. selective, faith, specialist, single-sex),

[2] A pseudo-candidate is a composition of different candidates sitting the same examinations from the same awarding body.

school size. This methodology could also be used to compare awarding bodies' results provided appropriate adjustments have been made. Any residual inter-awarding body differences after accounting for legitimate measures of ability might be thought to imply a difference in standards [6, 7].

4. Identifying an appropriate methodology

Each of the three strands described in *Section 3* addresses a number of key yet different issues.

For the **syllabus/specification review**, a comparison of content standards is made and the key questions addressed are:

- Are the demands of the syllabuses/specifications, examination papers & mark schemes similar?
- If they differ, how and where do they differ?
- What inferences can be drawn about the knowledge, understanding and skills of examinees who have taken these syllabuses?

Apart from identifying the differences in content standards between the two awarding bodies, the benefits of conducting such an exercise can usually inform examiners/setters about their styles[3] and language of question setting [8], depending on the granularity of how the research is being conducted. This review would also be useful to inform the update of syllabus specifications when the existing one has gone 'stale' after a period of time.

Cross-moderation belongs to the comparison of performance standards and the question addressed in here is:

- Which syllabuses' grade boundary scripts are perceived by expert judges to be of better quality (after allowing for the syllabus content, question paper and mark scheme difficulty)?

The concepts of Thurstone's pairs approach based on discriminal dispersion and the law of comparative judgement are used to compare paired sample scripts [1] in a cross-moderation methodology. The main advantage of this approach is that the use of candidates' scripts provides explicit evidences of the knowledge, understanding and skills of examinees who have taken these syllabuses, and hence direct comparison of performance standards can be achieved. It should be noted that it is only possible to compare performance standards if the content standards across the examination boards are similar enough for the different assessments to be considered to be measuring the same construct (underlying trait). If the question papers, mark schemes and syllabus specifications are very different, examiners will be expected to make judgements about the relative performance standards in a context of possible differences in content standards.

Statistical comparability belongs to the comparison of performance standards and the main question addressed is:

[3] The scaffolding and the Complexity, Resources, Abstractness, Strategy (CRAS) demand criteria.

- Are equivalent candidates (in terms of the control variables) equally likely to achieve a given grade in the examinations being compared?

As only statistical data are considered in this exercise, soft factors such as the improvement of teaching quality, change in teaching method, change in syllabus specifications will not be taken into consideration and hence a full extent of comparison might not emerge.

5. A rank-ordering methodology at syllabus level

The rank-ordering methodology at component level has been used relatively effectively to compare standards across boards among components. Previous research has found that rank-ordering generates a number of encouraging characteristics at component level including:

- comparability outcomes which resonate closely with other valid standard-maintaining activities in the UK assessment arena which use a range of statistical evidence in combination with expert judgment [2, 11];
- outcomes which are consistent over time; and
- a degree of flexibility in regard to experimental designs which allow:
    - examinations manifesting differentiated demands to be equated [11], and
    - post-hoc investigations of whether standards are being maintained over time across a number of consecutive sessions [10].

However, this approach becomes rather restricted when informing the adjustment of grading thresholds during awarding meetings at syllabus level and hence the rationale behind developing a methodology at syllabus level to tackle the issue. The modified methodology at syllabus level shares similar procedures of requiring consultants/examiners to rank-order candidates' scripts (and particularly pseudo-candidates scripts at syllabus level methodology), using multifacet Rasch analysis to analyse the data and using graphical representations to present research outcomes at component level. However, the script selection algorithm of pseudo-candidates and the evaluation pack design for consultants depart substantially from those at component level. Different aspects of the rank-ordering methodology at syllabus level [12, 13, 14] are described in each sub-section below.

5.1 Participants and materials

Five senior examiners/consultants preferably with marking/moderating experience of both syllabuses are recruited to evaluate pseudo-candidates scripts based on a holistic judgement. Their tasks are to rank-order scripts within each design pack from best (highest quality = 12) to worst (lowest quality = 1) and record their outcomes in the tables provided on a record sheet.

A researcher[4] is required to oversee the research and logistical aspects. The researcher is responsible for the design of the script selection algorithm, pack design relating to

[4] Also incorporating the responsibilities of a Project Manager.

the combination of pseudo-candidates' scripts, dealing with examiners queries during the evaluation period, data analysis and reporting. Project management responsibilities include the resources and project planning, managing administrative staff to carry out the logistical processes and making sure the entire project is running on time.

A member of administrative staff is required to carry out script cleaning to remove markings and annotations from examiners in case they influence rank-ordering decisions during evaluation. He/she is also required to carry out all logistical aspects relating to the project.

The research materials required in this project are question papers, mark schemes, syllabus specifications and candidates' scripts from both examination boards. FACETS (a multifacet Rasch software, version 3.64) was used in the data analysis and Microsoft Excel, with its potential for creative chart generation, was used to record and manipulate data prior to the analysis.

5.2 Procedures of conducting the rank-ordering research

The procedures described in this section provide brief guidelines for conducting the rank-ordering exercise [9].

A syllabus (Jun 2008)												
Core - option AA (3,11); Ext - option BB (4, 22)												
					wt = 104 to 104		wt = 130 to 130		wt = 56 to 56		wt = 70 to 70	
Grade	Within grade	Total option/syl level mark	Round-off aggregated total	Comp 3 (raw mk)		Comp 4 (raw mk)		Comp 11 (raw mk)		Comp 22 (raw mk)		Tier
A	A+2/3	148	148			101	101			47	47	Extended
A	A+1/3	139	139			96	96			43	43	Extended
A	A	129	129			90	90			39	39	Extended
A	A-1/3	119	119			82	82			37	37	Extended
A	A-2/3	110	110			77	77			33	33	Extended
A	B-1/3	90	90			63	63			27	27	Extended
C	C+2/3	90	90			63	63			27	27	Extended
C	C+1/3	81	81			58	58			23	23	Extended
C	C	71	71			51	51			20	20	Extended
C	C-1/3	67	67			48	48			19	19	Extended
C	C-2/3	62	62			45	45			17	17	Extended
C	D-1/3	54	54			39	39			15	15	Extended
E	E+2/3	54	54			39	39			15	15	Extended
E	E+1/3	49	49			35	35			14	14	Extended
E	E	45	45			32	32			13	13	Extended
E	E-1/3	41	41			29	29			12	12	Extended
E	E-2/3	36	36			25	25			11	11	Extended
E	F-1/3	NA	NA			NA	NA			NA	NA	Extended
C	C+2/3	111	111	68	68			43	43			Core
C	C+1/3	104	104	64	64			40	40			Core
C	C	98	98	60	60			38	38			Core
C	C-1/3	92	92	56	56			36	36			Core
C	C-2/3	85	85	52	52			33	33			Core
C	D-1/3	73	73	44	44			29	29			Core
E	E+2/3	73	73	44	44			29	29			Core
E	E+1/3	67	67	40	40			27	27			Core
E	E	61	61	36	36			25	25			Core
E	E-1/3	54	54	32	32			22	22			Core
E	E-2/3	47	47	27	27			20	20			Core
E	F-1/3	34	34	20	20			14	14			Core
Note: 3 scripts (even for repeated raw mks) are required from each highlighted yellow box for the specified raw mark.												

Figure 1. A script-pulling list of A SYLLABUS of exam board X. Grade thresholds of pseudo-candidates at and around Grades A, C and E were identified.

i) Identify the syllabus option (and hence components) between both exam boards by considering the largest entry number, the closest similarities in terms of syllabus specifications and nature/structure of the syllabus;

ii) construct script-pulling lists for key grade thresholds, e.g. A, C and E (each at 2/3 and 1/3 of a grade above the key grade, at the key grade, 1/3 and 2/3 of a grade below the key grade, and 1/3 of a grade below the next grade) for both boards as shown in Figure 1;

iii) identify 3 pseudo candidates' scripts at each grade threshold of the designated components for both boards and select 1 pseudo candidate's script out of the 3 based on legibility, script text (preferably written in black pen) and maximum question coverage within a paper;

iv) remove any examiner markings/annotations such that they do not have an influence on the rank-ordering decisions during examiners' evaluation process;

A SYLLABUS: PACK A

		Comp 4			Comp 22			
	Candidate	Centre No.	Candidate No.	Raw Mark	Centre No.	Candidate No.	Raw Mark	CODE
A+2/3	1	JP200	0060	101/130	QA026	0429	47/70	A3
A+1/3	2	ZM400	0048	96	QA026	0483	43	A2
A	3	JP200	0029	90	OM009	0074	39	A1
A-1/3	4	SA100	1005	82	OM009	0064	37	A6
A-2/3	5	BR100	0325	77	SA163	0015	33	A5
B-1/3	6	BR100	0304	63	ID062	6034	27	A4

The pseudo candidates are sorted using cand no of paper 4

	Paper 3 (Scoris)				Paper 4 (Scoris)			Paper 6			
	Candidate	Centre No.	Candidate No.	Raw Mark	Centre No.	Candidate No.	Raw Mark	Centre No.	Candidate No.	Raw Mark	CODE
A+2/3	7	61903	3049	77/100	12527	4078	70/100	20153	3083	41/48	A7
A+1/3	8	26338	8133	72	19227	9087	66	22161	3411	39	A9
A	9	20049	3212	66	16511	7036	60	12527	4159	37	A10
A-1/3	10	39255	4162	60	20833	3194	54	12460	3119	35	A8
A-2/3	11	23362	3099	55	22127	1511	50	19226	8124	33	A12
B-1/3	12	50735	4362	43	66539	3110	38	24324	5094	29	A11

The pseudo candidates are sorted using cand no of paper 6

Figure 2. A 'Grade A' pack design for exam board X (cands 1 – 6) and exam board Y (cands 7 – 12). Pseudo candidates were coded (code A1 to A12) to randomise the original rank order shown under the column 'Candidate'.

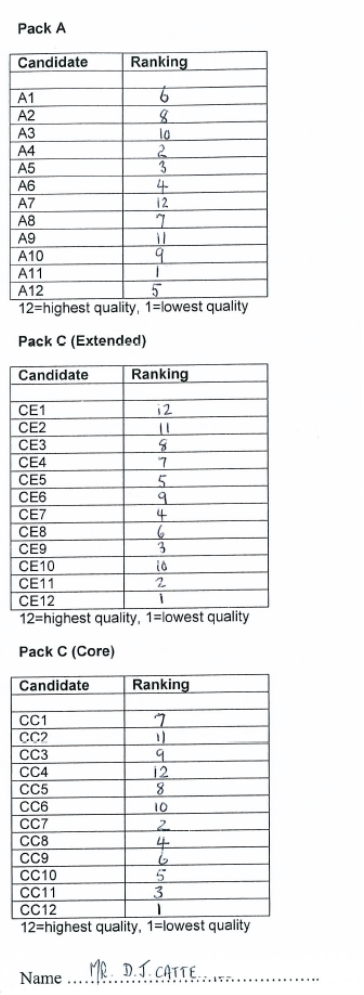
v) randomise, code and label the pseudo-candidates' scripts such that the original scripts' rank-order based on marks is concealed (see Figure 2);

vi) photocopy each script five times for each examiner/consultant;

vii) put the scripts into different design packs for each grade threshold according to the design pack lists, e.g. A, C and E, as shown in Figure 2 (also C Core, C Extended and C Core vs. Extended if either one of the exam board's syllabus is tiered);

viii) send instructions, rank-ordering score sheet, questionnaire and packs with scripts to Examiners for evaluation (typical evaluation period is two days). Examiners return the results to the examination board for data analysis.

5.3 Analysis and results



Pack A

Candidate	Ranking
A1	6
A2	8
A3	10
A4	2
A5	3
A6	4
A7	12
A8	7
A9	11
A10	9
A11	1
A12	5

12=highest quality, 1=lowest quality

Pack C (Extended)

Candidate	Ranking
CE1	12
CE2	11
CE3	8
CE4	7
CE5	5
CE6	9
CE7	4
CE8	6
CE9	3
CE10	10
CE11	2
CE12	1

12=highest quality, 1=lowest quality

Pack C (Core)

Candidate	Ranking
CC1	7
CC2	11
CC3	9
CC4	12
CC5	8
CC6	10
CC7	2
CC8	4
CC9	6
CC10	5
CC11	3
CC12	1

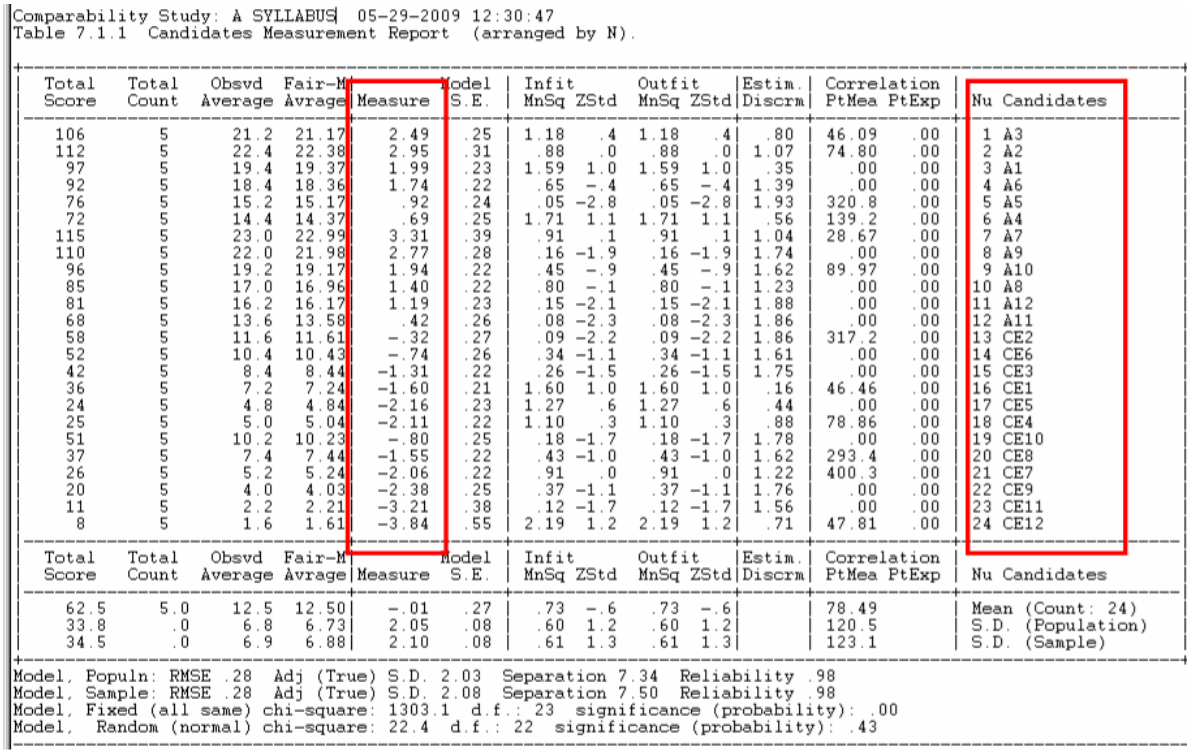
12=highest quality, 1=lowest quality

Name ..... MR D.J CATTE .....

Figure 3. An examiner's rank-ordering results for design packs A, C(Ext) and C(Core).

Each examiner generated rank-order data. A typical example of the data is shown in Figure 3. Figure 3 shows the results for design packs A, C (Ext) and C (Core), but not for E. The rank-order data together with the syllabus percentage mark at each corresponding grade threshold was then re-arranged in Excel to comply with the FACETS format before inputting into the program. It should be noted that the percentage mark, instead of a raw mark, at syllabus level is used in the analysis in order to achieve a common scale for both examination boards; and two facets, *Rater* and *Script*, were used in the program. Figure 4 shows the *Candidates Measurement Report* routinely generated by FACETS. Columns of *'Measure' (script quality)*, *'Nu'* and *'Scripts'* highlighted in red are used. Figure 5 illustrates an example of the comparability plot with accompanying commentary at Grade A and Grade C (Extended option).

In Figure 5, the vertical axis along the left of the figure represents the 'Measure' (or script quality) scale. This scale is common to both person ability and script difficulty



Comparability Study: A SYLLABUS 05-29-2009 12:30:47
Table 7.1.1 Candidates Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu	Candidates
106	5	21.2	21.17	2.49	.25	1.18	.4	1.18	.4	.80	46.09	.00	1	A3
112	5	22.4	22.38	2.95	.31	.88	.0	.88	.0	1.07	74.80	.00	2	A2
97	5	19.4	19.37	1.99	.23	1.59	1.0	1.59	1.0	.35	.00	.00	3	A1
92	5	18.4	18.36	1.74	.22	.65	-.4	.65	-.4	1.39	.00	.00	4	A6
76	5	15.2	15.17	.92	.24	.05	-2.8	.05	-2.8	1.93	320.8	.00	5	A5
72	5	14.4	14.37	.69	.25	1.71	1.1	1.71	1.1	.56	139.2	.00	6	A4
115	5	23.0	22.99	3.31	.39	.91	.1	.91	.1	1.04	28.67	.00	7	A7
110	5	22.0	21.98	2.77	.28	.16	-1.9	.16	-1.9	1.74	.00	.00	8	A9
96	5	19.2	19.17	1.94	.22	.45	-.9	.45	-.9	1.62	89.97	.00	9	A10
85	5	17.0	16.96	1.40	.22	.80	-.1	.80	-.1	1.23	.00	.00	10	A8
81	5	16.2	16.17	1.19	.23	.15	-2.1	.15	-2.1	1.88	.00	.00	11	A12
68	5	13.6	13.58	.42	.26	.08	-2.3	.08	-2.3	1.86	.00	.00	12	A11
58	5	11.6	11.61	-.32	.27	.09	-2.2	.09	-2.2	1.86	317.2	.00	13	CE2
52	5	10.4	10.43	-.74	.26	.34	-1.1	.34	-1.1	1.61	.00	.00	14	CE6
42	5	8.4	8.44	-1.31	.22	.26	-1.5	.26	-1.5	1.75	.00	.00	15	CE3
36	5	7.2	7.24	-1.60	.21	1.60	1.0	1.60	1.0	.16	46.46	.00	16	CE1
24	5	4.8	4.84	-2.16	.23	1.27	.6	1.27	.6	.44	.00	.00	17	CE5
25	5	5.0	5.04	-2.11	.22	1.10	.3	1.10	.3	.88	78.86	.00	18	CE4
51	5	10.2	10.23	-.80	.25	.18	-1.7	.18	-1.7	1.78	.00	.00	19	CE10
37	5	7.4	7.44	-1.55	.22	.43	-1.0	.43	-1.0	1.62	293.4	.00	20	CE8
26	5	5.2	5.24	-2.06	.22	.91	.0	.91	.0	1.22	400.3	.00	21	CE7
20	5	4.0	4.03	-2.38	.25	.37	-1.1	.37	-1.1	1.76	.00	.00	22	CE9
11	5	2.2	2.21	-3.21	.38	.12	-1.7	.12	-1.7	1.56	.00	.00	23	CE11
8	5	1.6	1.61	-3.84	.55	2.19	1.2	2.19	1.2	.71	47.81	.00	24	CE12

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu	Candidates
62.5	5.0	12.5	12.50	-.01	.27	.73	-.6	.73	-.6		78.49			Mean (Count: 24)
33.8	.0	6.8	6.73	2.05	.08	.60	1.2	.60	1.2		120.5			S.D. (Population)
34.5	.0	6.9	6.88	2.10	.08	.61	1.3	.61	1.3		123.1			S.D. (Sample)

Model, Populn: RMSE .28 Adj (True) S.D. 2.03 Separation 7.34 Reliability .98
Model, Sample: RMSE .28 Adj (True) S.D. 2.08 Separation 7.50 Reliability .98
Model, Fixed (all same) chi-square: 1303.1 d.f.: 23 significance (probability): .00
Model, Random (normal) chi-square: 22.4 d.f.: 22 significance (probability): .43

Figure 4. The relevant FACETS output of A SYLLABUS. The columns of Measure, Nu and Scripts were used for the comparability plot in Figure 5.

and is measured in logits (or log odds)[5] . In Rasch measurement, the logit is a way of expressing the probability (or odds) of a particular event. The 'Measure' scale is an equal interval scale, that is, it can tell us not only that one script[6] is more *difficult* than another, but also how much more difficult it is. The equal interval nature of the *ability* measurements means that growth in ability over time can be plotted on the scale.

In these graphs each data point (square - Board X and diamond - Board Y) represents a 'script'. Each script (a data point) is positioned according to its 'Measure'. For example, a score of +2 logits indicates a higher quality script than a score of –2 logits. Thus performances are rank ordered with the most able candidates at the top of the axis and the least able at the bottom, that is, the scripts in the top half of the graph (above 0 logits) are judged to be of better quality than those in the bottom half (below 0 logits).[7]

The horizontal axis shows the overall syllabus percentage.

[5] The mathematical unit of Rasch measurement, the log-odds unit or 'logit', is defined prior to the experiment. One logit is the distance along the line of the variable that increases the odds of observing the event specified in the measurement model by a factor of 2.718.., the value of 'e', the base of 'natural' or Napierian logarithms used for the calculation of 'log-' odds. All logits are the same length with respect to this change in the odds of observing the indicative event.

[6] A script is defined here as a set of performances (a pseudo-candidate) representing components within an option.

[7] In Rasch terms, a rank ordering looks like the outcome of a one-facet test. The performances of examinees (scripts) are compared with each other, either by direct encounter or by the examiners's thought-experiments. Thus the final ordering no longer has any quantifiable connection with the difficulty of the elements of performance on which the comparison was made, or the severity of the examiners who constructed the orderings. Removing examiner severity and script difficulty from consideration is often an intended aim of rank ordering. Whilst this type of data does not appear to be amenable to the familiar axioms of fundamental measurement, objective measurement is possible with rank ordered data.

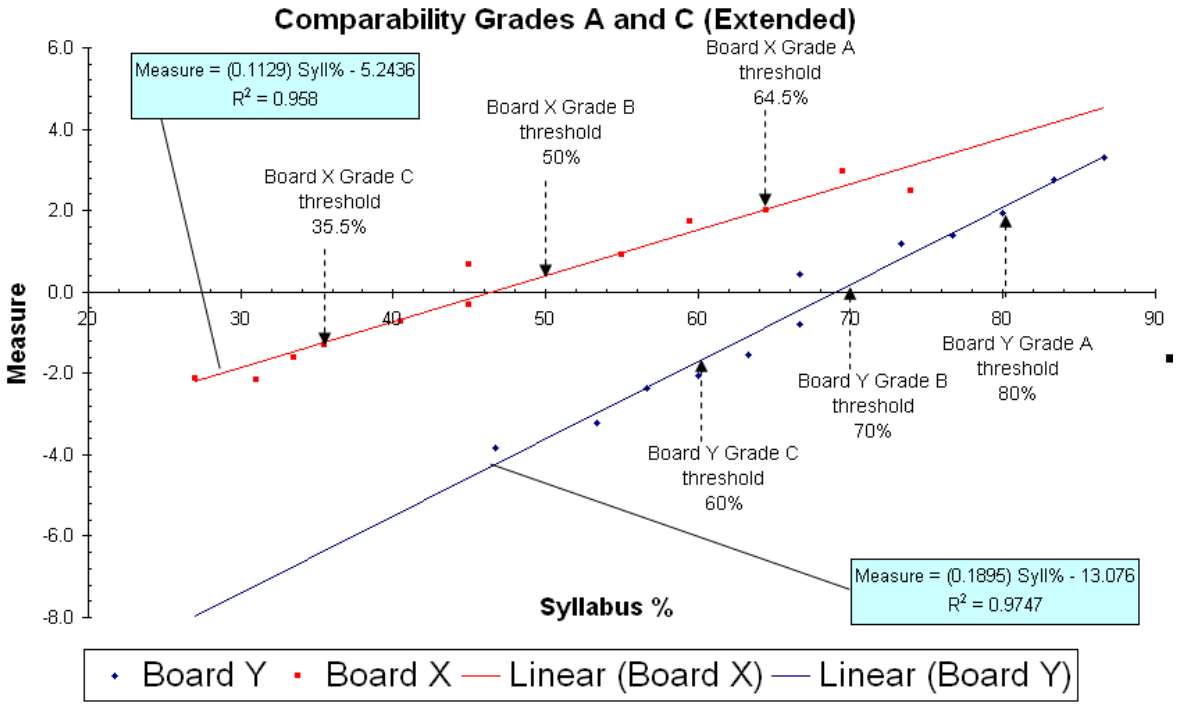


Figure 5. A comparability plot of A SYLLABUS grades A and C (Extended option) between exam boards X and Y. Board X is equivalent to Board Y at A, but severe at B and C (Ext).

The red (Board X) and blue (Board Y) lines in the example graph shown in Figure 5 are linear regression lines and the regression equations are illustrated in the blue boxes. The parameter R in the blue box is the correlation coefficient between the variables *Measure* and *Syllabus %*. Interpreting a correlation coefficient is possible if it is perceived in terms of the overlap between the two variables. In order to observe the overlap, the square of the correlation coefficient must be taken so that it is possible to see how much of the variance in one measure can be accounted for by the other measure and hence the parameter R^2.

The interpretation of Figure 5 at Grades A, B and C (Extended option shown as an example) is described as follows.

Grade A: A candidate with an ability measure of +2.04 logits would just be able to scrape an A grade (Board X) but would need an infinitesimally higher ability to achieve the same grade on Board Y. The Board Y threshold was set so that a candidate with an ability of +2.08 logits just achieved a Grade A. In order for a candidate with the same ability to achieve a Board Y Grade A, Board X would have needed to set their A grade threshold at 64.9% (instead of 64.5%). As things stand, the Board X and Board Y thresholds were correctly set for this examination session.

Grade B: A candidate with an ability measure of +0.19 logits would just be able to scrape a B grade (Board Y) but would need a higher ability to achieve the same grade on Board X. The Board X threshold was set so that a candidate with an ability of +0.40 logits just achieved a Grade B. In order for a candidate with the same ability to achieve a Board Y Grade B, Board X would have needed to set their B grade

threshold at 48.12% (instead of 50%). As things stand, the Board X threshold was severe by approximately 1.88% of the syllabus marks.

Grade C (Extended): A candidate with an ability measure of -1.71 logits would just be able to scrape a C grade (Board Y) but would need a higher ability to achieve the same grade on Board X. The Board X threshold was set so that a candidate with an ability of -1.24 logits just achieved a Grade C. In order for a candidate with the same ability to achieve a Board Y Grade C, Board X would have needed to set their C grade threshold at 31.33% (instead of 35.5%). As things stand, the Board X threshold was severe by approximately 4.17% of the syllabus marks.

The percentage difference of syllabus marks at grade thresholds A, B and C between the two boards provides further information to inform the awarding committee to make the threshold adjustment decision together with other grading information.

6. Feedback from examiners during the evaluation

Questionnaire responses were collected from examiners to help understand the qualitative aspects of the study relating to: the overall difficulty of the task, the amount of time taken to rank order the scripts, what made some packs more or less difficult to rank, the difficulties presented by 'pseudo/composite' candidates; any differences in the task between papers, and the strategy they deployed. The responses presented here are summaries from several comparability studies.

6.1 Overall difficulty of the task

The majority of examiners found the task either 'fairly difficult' or 'very difficult' to execute, only a few examiners felt that they were reasonably comfortable with the activity and claimed that once a strategy had been formulated the exercise was comparatively straightforward. Reasons for difficulty included:

- initial difficulties engendered by the enormity of the task which eased once a 'method' had been established
- unfamiliarity with syllabuses
- disparities in syllabus and mark scheme demand/composition
- inconsistent quality across pseudo/composite candidates' profile
- disparate skill sets across papers
- discrepancies between question papers
- retention of marking criteria across papers
- resisting tendency to remark
- first time experience of rank ordering

Examiners tended to take between 60 and 240 minutes per pack during the evaluation. A few examiners claimed to spend approximately 40 minutes per pack.

Differences were also reported relating to the ease or difficulty of rank ordering certain packs. Extended packs were the most time-consuming to rank order although Grade A packs were slightly less problematic as there was a wider range of ability instantiated in performances. Scripts from less able candidates were more difficult to rank, and standards were more closely grouped. Inexperience appeared to be the main

contributing factor to difficulty. Other factors included topic variation (student strengths being topic-related) and disparity in syllabus requirements.

Some examiners suggested that the task would have been made easier if they had ranked individual scripts from the same candidate. Other examiners articulated a range of difficulties associated with the nature of a 'pseudo/composite' profile:

- inauthentic performance profile
- 'pseudo-candidates' invariably demonstrate differing strengths, real candidates might give more clues along the way [15, 16]
- evidence of different pedagogical heritage

6.2 Rank–ordering strategy

Examiners were allowed to adopt their own rank-ordering strategy during the evaluation phase though they were not allowed to re-mark the scripts. A variety of strategies were identified:

- employ an item level analysis using ticks and crosses (a numeric analysis would provide a better overall judgement)
- multiple readings of pairs of scripts with brief notes which inform successive re-definitions
- syllabus comparative analysis – determining which parts of the syllabus have been mastered
- experience facilitates identification of questions which are poorly attempted by less able students and which questions are well answered
- a point awarding system for each paper given the diversity of performance
- judging first what is expected of an average (C grade) candidate and using as a comparator
- provisional order followed by multiple re-reads. This is accomplished in batches (good; medium; poor candidates) and then by group
- judgment based on individual questions
- initial question answering followed by mark scheme verification. Re-marking of scripts, batch-by-batch, question-by-question. The whole process being subject to fine-tuning
- identifying questions indicative of student ability and matching these with comparable questions across papers
- question ranking: A – F, on each set of scripts, providing an overall grade for separate scripts. Correct use of mathematical formulae signalling individual ability levels
- custom-made, inspection-determined scoring system. Scripts with tied scores are subject to closer scrutiny

The majority of examiners indicated a change of approach as the rank order task became increasingly more familiar. With experience, greater confidence was placed in subsequent judgements; fewer notes were made; and a greater tendency to revisit and overturn earlier judgements was also reported [17].

Examiners were uncertain as to whether more or less time on each script made any difference to the final rank order. However, in the main, they believed that a reduction

or extension in the time taken to undertake the exercise would have little impact on the outcome.

7. Conclusions

A comparison of a syllabus between two examination boards based on the rank-ordering methodology at syllabus level has been carried out successfully. This methodology is a modified approach of similar rank-ordering methodologies at component level. One advantage of conducting rank-ordering comparability at syllabus level is that graphical outputs generated by multifacet Rasch indicate the difference in standards between two examination boards in terms of syllabus percentage which can inform threshold adjustment during awarding meetings. Feedback from examiners was also captured to help understand the evaluation process. Although the majority of examiners found the task either 'fairly difficult' or 'very difficult' to execute, almost all analyses and examiners' feedback showed that examiners were capable of completing the tasks. Examiners employed several different strategies to carry out the initial rank ordering exercise. As the evaluation progressed, the majority of examiners indicated a change of approach reflecting their familiarity of, and confidence in, the exercise. Areas of research relating to the effect of real candidate scripts versus pseudo/composite candidate scripts, and the effect of different script design/selection algorithms on the final rank-ordering outcome at syllabus level will constitute further areas for consideration.

Bibliography

1. Bramley, T. (2007) "Chapter 7. Paired Comparison methods". In Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (Eds.) Techniques for monitoring the comparability of examination standards, *Qualifications and Curriculum Authority*
2. Bramley, T.: "A rank-ordering method for equating tests by expert judgement", *Journal of Applied Measurement*, **6**(2), 202-223, 2005.
3. Newton, P (2007) "Chapter 1. Contextualising the comparability of examination standards". In Newton, P., Baird, J., Goldstein, G., Patrick, H and Tymms, P (Eds.) Techniques for monitoring the comparability of examination standards. *Qualifications and Curriculum Authority*
4. Baird, J (2007) *Alternative conceptions of comparability*. In Newton, P., Baird, J., Goldstein, G., Patrick, H and Tymms, P (Eds.) Techniques for monitoring the comparability of examination standards. Malta: Qualifications and Curriculum Authority
5. Hambleton, R.K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods and Perspectives.* (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum Associates.
6. Cresswell, M. J (1996) Defining, setting and maintaining standards in curriculum-examinations: judgemental and statistical approaches. In Goldstein, H and Lewis, T (Eds), *Assessment: Problems, Developments and Statistical Issues.* Chichester: John Wiley and Sons Ltd.
7. Schagen, I and Hutchison, D (2003) Adding Value in Educational Research – the marriage of data and analytical power. *British Educational Research Journal*. Vol. 29, No. 5
8. Evaluation and Psychometrics team, Research Division, ARD: 'Combined Science comparability study', *Cambridge Assessment Internal report*, July 2009.
9. Yim, L.W.K., Shaw, S.D. and Lewis, M.: 'A science comparability study between two exam boards using a rank-ordering methodology at syllabus level', *9th AEA Europe Conference Proceeding, Hisar, Bulgaria*, 6-8th Nov 2008, pp. 35.
10. Black, B (2007). Investigating January versus June awarding standards using an adapted rank-ordering method. *Cambridge Assessment Internal Report.*

11. Gill, T. and Black, B (2006). An investigation of standard maintaining and equating using expert judgment in GCSE English between years and across tiers using a rank-ordering method. *Cambridge Assessment Internal Report.*
12 Yim, L., Shaw, S. D and Batten, P (2008a) A Phase 2 Report on the Comparison of Physics between two exam boards – June 2007 session. *Cambridge International Examinations Internal Report.*
13. Yim, L., Shaw, S. D and Batten, P (2008b) A Phase 2 Report on the Comparison of Chemistry between two exam boards – June 2007 session. *Cambridge International Examinations Internal Report.*
14. Yim, L., Shaw, S. D and Batten, P (2008c) A Phase 2 Report on the Comparison of Biology between two exam boards – June 2007 session. *Cambridge International Examinations Internal Report.*
15. Arlett, S (2002) A Study in VCE Health and Social Care, Units 1, 2 and 5. A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2001 examination and organised by AQA on behalf of the Joint Council for General Qualifications.
16. Guthrie, K (2003) A comparability study in GCE Business Studies, Units 4, 5 and 6 VCE Business, units 4, 5 and 6. A review of the examination requirements and a report on the cross-moderation exercise. A study based on the summer 2002 examination. Organised by Edexcel on behalf of the Joint Council for general Qualifications.
17. Jones, B., Meadows, M. and Al-Bayatti, M (2004) Report of the inter-awarding body comparability study of GCSE religious studies (full course) summer 2003. Assessment and Qualifications Alliance.