



CAMBRIDGE ASSESSMENT

***Using an adapted rank-ordering method to investigate
January versus June awarding standards.***

Beth Black

Research Division
Cambridge Assessment
1 Regent Street
Cambridge
CB2 1GG
United Kingdom

Black.B@cambridgeassessment.org.uk

A paper presented at the Fourth Biennial EARLI/Northumbria
Assessment Conference, Berlin, Germany, August 2008.

<i>Using an adapted rank-ordering method to investigate January versus June awarding standards.</i>	1
Abstract	3
1. Introduction	4
2. Method	7
2.1 Selection of units	7
2.2 Obtaining the archives/the scripts	8
2.3 The judges.....	9
2.4 Pack design	9
2.4.1 Packs of three	9
2.5 Procedure.....	11
2.6 Analysis	11
3. Results	12
3.1 Analysis of Rank Order outcomes	12
3.2 Ranking question paper difficulty	14
3.3 Judge feedback on the activity	16
3.3.1 Difficulty of the overall task.....	16
3.3.2 Allowing for differences in question paper difficulty	16
3.3.3 Were some scripts easier to rank than others?	17
3.3.4 Rank ordering strategies.....	18
3.3.5 Differences and similarities between making judgements in the context of rank-ordering versus awarding.	19
4. Discussion.....	20
4.1 Limitations of this research.....	20
4.2 Other issues	20
Appendix A: Rank Ordering Instructions.....	23
Appendix B: Question Paper ranking activity	24
Appendix C: Example Feedback Questionnaire	25
Appendix D Tables of One-way ANOVA results.	27
Appendix E Tables of Two-way ANOVA results	28

Abstract

The dual aims of this research were (i) to pilot an adapted rank-ordering method and (ii) to investigate whether UK examination awarding standards diverge between January and June sessions.

Standard maintaining is of critical importance in UK qualifications, given the current 'high stakes' environment. At qualification level, standard maintaining procedures are designed to ensure that a grade A in any particular subject in one year is comparable to a grade A in another year, through establishing the equivalent cut-scores on later versions of an examination which carry over the performance standards from the earlier version.

However, the current UK method for standard setting and maintaining - the awarding meeting as mandated by the QCA Code of Practice (2007) - introduces the potential for the awarding standards of the January and June sessions to become disconnected from each other. Additionally, from a regulatory and comparability research point of view, the January sessions have been largely ignored, despite the increasing popularity of entering candidates for January units since Curriculum 2000.

Given the difficulties in quantifying relevant features of the respective cohorts, (e.g. the January candidature is more unstable), and the problems in meeting the assumptions necessary for statistical methods (e.g. Schagen and Hutchison, 2008), arguably the best way to approach this research question is to use a judgemental method focusing on performance standards. In this study, the chosen method involves expert judges making comparisons of actual exemplars of student work ('scripts').

A rank-order method was employed adapted from Bramley (2005). Archive scripts at the key grade boundaries (A and E) from the previous six sessions (comprising three January and three June sessions) from two AS level units in different subjects were obtained.

Whilst previous rank order exercises (e.g. Black and Bramley, in press) required judges to rank order ten scripts per pack spanning a range of performance standard, in this study each exercise involved scripts which were, at least notionally, of more similar quality (e.g. all exactly E grade borderline scripts) and therefore an adaptation was required. Rank-ordering sets of three scripts retains many of the advantages of rank-order over traditional Thurstone paired-comparisons as well as an additional advantage - asking of judges a more natural psychological task: to simply identify, on the basis of a holistic judgement, the best, middle and worst script.

Rasch analysis of the rank order outcomes produced a measure of script quality for each script and, using an ANOVA, it was possible to examine effects of session and session type (i.e. January versus June). The research indicated that the two AS units displayed different patterns of performance standards for January versus June.

The discussion will be research-related and practice-related. The paper will address the potential for using this method to investigate comparability in a variety of contexts, and implications for standard-maintaining processes.

1. Introduction

Standard maintaining is of critical importance in UK qualifications, given the 'high stakes' environment in which they function. At qualification level, standard maintaining procedures are in place to ensure that a grade A in Physics in one year is of comparable demand to a grade A another year, or even several years later. Essentially, standard maintaining involves establishing the equivalent marks on later versions of an examination which carry over the performance standards from the earlier version. That standards are maintained is a fundamental premise of the Universities and Colleges Admissions Service's (UCAS) tariff system, where A level grades are converted into currency to gain admissions to university. The currency gained from an A grade in one year, in the UCAS tariff, is identical to the currency gained from an A grade in another year. Thus, comparability of standards over time and across sessions (as well as between subjects) is vital for fairness – that candidates from one session should not be advantaged or disadvantaged by awarding decisions in comparison to other sessions of the same examination, within a reasonable and relevant time frame.

Most UK comparability work to date has focused upon the main examination session in a year, namely the summer session i.e. when the bulk of GCSE, AS and A level candidates sit examinations and aggregate towards qualifications. For example, the QCA Standards Over Time series¹ has entirely sourced question papers and script evidence from the summer sessions. From a regulatory and comparability research point of view, the January sessions have been largely ignored. This is despite the increasing popularity of entering candidates for January units since the introduction of Curriculum 2000 spawning an increase in modular specifications. The growing popularity of January entry is due to not only opportunity (which on linear specifications this did not exist), but also convenience: notwithstanding requirements for synopticity, January entry allows students to revise, parcel up and leave behind one slab of content, reducing the so-called 'assessment burden' for students in the summer series; additionally, January entry affords students more scope for retaking units (which may be of particular importance for A2 candidates keen to meet their UCAS offer). It is thus vital to ensure that achievement at any particular grade should not be more or less demanding in January than June.

Currently, the current method for standard setting and maintaining is the awarding meeting, as mandated by the Qualifications and Curriculum Authority (QCA) Code of Practice (2008). This involves deciding upon a mark for each key grade boundary² through a consideration of a blend of both statistical and judgemental information. Firstly, the statistical information reviewed by the committee involves comparisons with previous sessions in terms of raw score distributions, cumulative pass rates and information about the cohort (such as centre type, putative grades³, forecast grades⁴). Secondly, the judgemental aspect of awarding involves a committee of senior examiners. Once a range of marks has been determined by the Principal Examiner (PE) and Chair of Examiners⁵ as broadly where the boundary will lie,

¹ Standards Over Time reports can be accessed at http://www.qca.org.uk/12086_1509.html

² For A level units, the key grade boundaries are the A/B E/U grade boundaries. Intervening grades are calculated arithmetically.

³ Putative grades are predictions of grade distributions for any specification or unit. They are based upon known data about the cohort in terms of prior attainment (i.e. GCSE grades). In a sense, they constitute population adjusted cohort-referencing, as putatives are not based upon actual candidate performance in the unit in question.

⁴ Frequently regarded as the least reliable indicator, these are the teachers' forecasts at grade level of a student's performance.

⁵ The PE is the person appointed by the awarding to set the question paper and lead the standardisation of the marking. The Chair of Examiners is responsible to the awarding body for

the senior examiners are asked to judge script evidence at all mark points in this range. The judgemental task is essentially binary in that the script is judged as worthy or not worthy of the higher grade of the pair (e.g. worthy of an A grade rather than a B grade). In order to guide the professional judgement and ensure maintenance of standard year on year, session on session, the senior examiners should use archive scripts as an anchor point and refer to performance descriptors where available. Archive scripts used for this purpose have always been sourced from the previous June session (in practice to date but in process of change).

Determining grade boundaries for a January may be considered more challenging. The size of the entry, as well as its composition in terms of centre type, might fluctuate more from one session to the next than it does in June. Additionally, candidates are likely to be more diverse in January: many will be 4-5 months younger than the June cohort and had less overall exposure to the subject; but many will be re-takers and, while perhaps not the highest ability candidates, may have benefited from more exposure to the subject. Possibly because of this, the statistical reference points (once a specification is beyond its second awarding session) tend to be mainly the previous 'like' session e.g. June 2006 data will be the main reference point for June 2007 awarding; not January 2007. This practice is common across all unitary awarding bodies. Certainly, the QCA Code of Practice (2008) introduces the potential for the awarding standards of the January and June sessions to become disconnected from each other. The injunction for the awarding committee to consider "information on candidates' performance in at least the two previous *equivalent* series" (my italics) (ibid 6.15 viii) means that continuity of key statistical indicators is required only from June to June and January to January and not necessarily between the two series.

Arguably, then, such standard maintaining practices introduce scope for the standards between January and June sessions to diverge, as suggested by figure 1. This is the hypothesis which this research project intended to investigate.

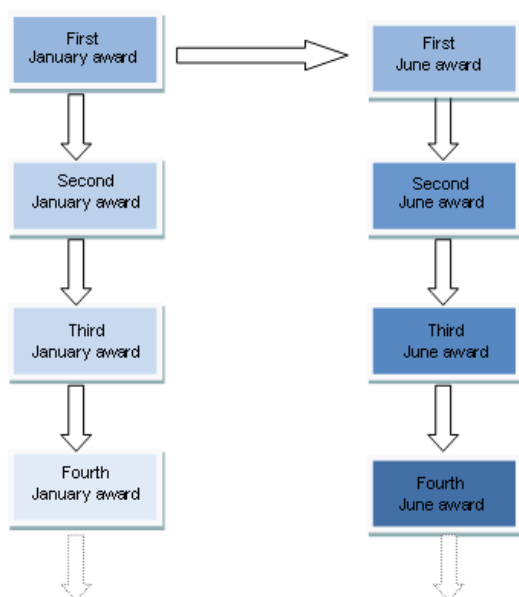


Figure 1: Diagram representing how statistical information is used to inform awarding decisions. From the third award onwards (in this example, this would be the second January award), the reference statistics are from the previous 'like' session i.e. January to January and June to June.

maintaining standards from year to year. The roles and responsibilities of awarding body personnel are fully set out in the QCA Code of Practice (2007)

The main research question of this study is whether there is any evidence that January and June sessions do diverge with regard to standards at key grade boundaries. Given the difficulty described above in quantifying relevant features of the cohorts, and the problems in meeting the assumptions which would be needed for statistical methods (e.g. Newton (1997); Baird (2007); Schagen & Hutchison (2007)), arguably the best way to approach this question is to use a judgemental method, involving comparisons of actual exemplars of student work and expert judges who are fully conversant with the content of the specification, the demands of the question papers and the performance standards required at the grade boundaries.

Opinion or script scrutiny work were rejected as methods for this study as lacking rigour and formality. Instead, an adapted rank order method was selected. Previous research (Black and Bramley, 2008; Bramley and Black, 2008 and Bramley, 2005, Bramley, Gill and Black 2008) gives much evidence and confidence that is valid method for conducting comparability exercises. In essence, a rank-order tasks involves judges sorting scripts in order of perceived quality based upon a holistic judgement.

Previous rank order exercises (Black and Bramley, 2008; Bramley 2005; Gill, Black and Bramley in prep) had utilised packs of ten scripts each. Such exercises had required judges to rank order ten scripts from best to worst. This was considered to be a reasonably achievable task in the context of these studies where each of the packs spanned a significant segment of the mark range. Additionally, feedback from judges in these studies indicated that while the task was not necessarily easy and while the cognitive demand of retaining judgements of script quality for ten scripts simultaneously was not inconsiderable, it was certainly do-able.

However, in this study, because, at least notionally, all scripts should be of the same quality (e.g. all exactly E grade borderline scripts) it was decided that requiring judges to rank order ten scripts would either be overly cognitive demanding or simply an impossible task. While a simple Thurstone paired design (as used for Joint Awarding Body comparability exercises and QCA's Standards Over Time series) was considered for this task, we decided instead to use packs of three (which we will call 'Thurstone Triples'), which have certain advantages in this context:

- It is still a cognitively meaningful task (as judges decide upon a winner, a loser and one in the middle i.e. they are ranked top, middle and bottom).
- Packs of three produce three 'paired comparisons' (see figure 2). So in comparison to a standard Thurstone pairs design, this triples the number of comparisons generated with less than triple the time or cognitive expenditure. Thus, this is a more efficient way to maximise the number of comparisons generated for the whole exercise.⁶
- They offer a more versatile design in that scripts in one pack or set of packs can, if desired, be presented and compared from more than two sessions (or specifications) in a more seamless fashion.

⁶ Any method which involves the use of scripts is necessarily time-intensive (the scripts must be read) and therefore expensive. Thus, studies involving the use of script-based evidence are carefully planned to maximise data and minimise judge time and effort,

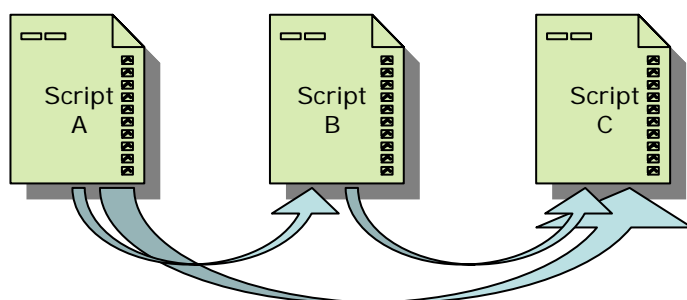


Figure 2: Comparisons generated from rank ordering three scripts: AB, BC and AC.

It was thought that using Thurstone Triples, as a method of choice over Thurstone pairs, might also ensure that judges could be less prone to demand characteristics (or overt 'fixing') in that judges have more variably constructed packs and are not repeatedly making binary comparisons.

Section 2 of this report describes the methodological aspects of the study (including choice of units, scripts used, design of packs, the judges and the procedure). Section 3 describes the results of the rank order exercise, the judges' rank ordering of question paper difficulty (rather than script performance), and the qualitative feedback from the judges including their perception of the level of difficulty of the exercise and the strategies they used. The report concludes with a discussion including suggestions for further research, further applications of the Thurstone triples technique and a discussion of the limitations of this study.

2. Method

2.1 Selection of units

The two units that were selected were both from AS qualifications:

- Economics AS unit 2882 ('Market Failure and Government Intervention') and
- Psychology AS unit 2540 ('Core Studies 1').

While there was an opportunist element regarding the selection of units (i.e. units for which the necessary scripts could be obtained), they met three criteria relating to stability features:

- The specification had remained essentially unchanged since 2000 and the question paper format had remained unchanged since June 2003 (when the three hour rule⁷ was introduced).
- Both units have a candidature in excess of 1500 for the January session, thus avoiding any of the problems which may be encountered through awarding small entry units.
- Both units have a relatively stable cohort in each session (January on January; June on June) over the period in question (see figure 3).

Psychology Unit 2540 consists of 20 short answer questions each worth between 2 and 4 marks.

Economics Unit 2882 consists of 3 or 4 structured or multi-part questions (short and medium) answer questions each worth between 2 and 6 marks as well as 2 questions requiring more extended writing with a 10 to 12 mark tariff.

⁷ The "three hour rule" was introduced as a response to some of the remonstrations surrounding 'over-assessment'. As a result, many AS units were reduced in examined time (and therefore format) so that the total AS examined time did not exceed 3 hours.

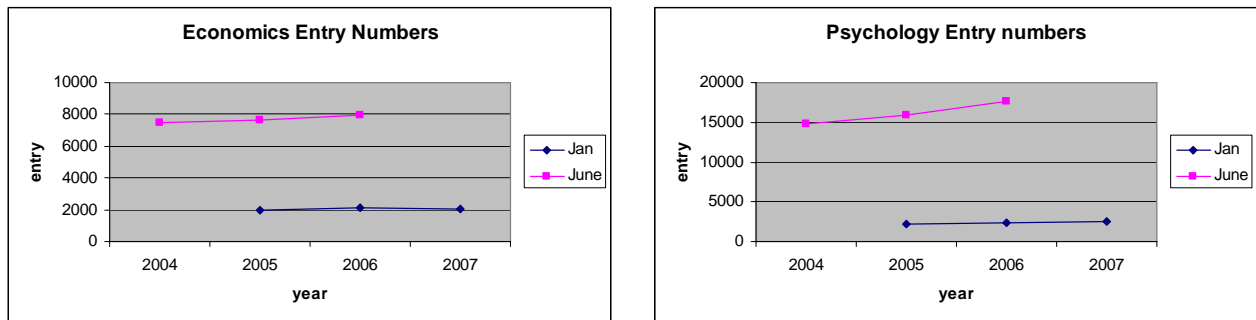


Figure 3: Candidature for Economics and Psychology units between June 2004 and January 2007.

2.2 Obtaining the archives/the scripts

For many reasons, awarding bodies retain few scripts once an examination session is over. Thus, this research did not have the luxury of a free hand in obtaining scripts from any previous examination series. However, compliance with the Code of Practice by maintaining a script archive from each session (para 6.30) meant that a limited amount of archive material was available:

While a specification remains in use, the awarding body must maintain a full archive containing marked scripts at each key grade boundary covering both series for at least the last five series. In addition, the awarding body must retain equivalent evidence from the first examination of the specification to guide the work of examiners and awarders.

As mentioned above, the purpose of the archive scripts is that they should be used in Awarding (as a requirement of the Code of Practice) in order for the committee to judgementally match the live session with the prior session (para 6.15 iii). To best further this end, scripts selected for archive should be illustrative of the standard at the key grade in question. They are usually selected by the Principal Examiner from those live scripts available at the Awarding meeting (and these scripts tend to have been marked by more experienced and senior examiners). As well as legibility perhaps being a criterion, Principal Examiners try to ensure that the scripts selected for archive embody the standard of a borderline A or E grade candidate. Therefore, it could be argued that archive scripts are more representative of the key grade boundary mark than a randomly obtained sample on that mark point.

The archive for these two units held five scripts for each of the previous five sessions⁸. A further five-script archive was obtained from the session immediately prior to the research activity (January 2007) in order to balance the number of January and June scripts. Thus, for each key grade boundary and for each unit, 30 scripts were obtained in total, comprising five each from the following six consecutive sessions:

- June 2004
- January 2005
- June 2005
- January 2006
- June 2006
- January 2007

⁸ June 2004, June 2005, June 2006, January 2005, January 2006

Each of these scripts was electronically scanned. The digital image was cleaned of marking annotation including all numerical marks, all qualitative comments and as many ticks as was feasible. The scripts were then reprinted prior to being placed in a pack.

2.3 The judges

Twelve judges were recruited in total, six for Economics AS unit 2882 and six for Psychology AS unit 2540. For the Economics, the judges were all Principal Examiners for the Economics specification (3812/7812) and had all been awarding committee members. For Psychology, of the six judges, five were Principal Examiners for the AS/A level specification (3876/7876) and one a former Team Leader who had all successfully taken part in a previous rank order exercise (Black and Bramley, 2008).

2.4 Pack design

2.4.1 Packs of three

As noted earlier, rank ordering script triples was considered the most apposite design for this task for reasons of the balance of efficiency of producing paired comparison data compared to cognitive expenditure, the naturalness of the task, and the likely validity and integrity of the comparisons.

2.4.2 Overall pack design

Each of the six judges had 30 packs: 15 packs for the E grade boundary scripts and 15 packs for the A grade boundary scripts. Each pack comprised three scripts.

The guiding principles for the overall pack design of the study were:

- To maximise each judge's exposure to the available range of scripts by aiming for each judge to see each available script at least once (i.e. for each judge to see all 30 scripts from the key grade boundary in question at least once).
- To minimise the number of times a judge saw any one script. (No judge saw any script more than twice).
- To ensure that no single judge made the same paired comparison twice.
- For each judge to have a unique set of packs and combinations of scripts.
- To maximise the total number of unique paired comparisons generated by the exercise. For each grade boundary, having obtained 30 scripts for each unit, the total number of unique paired comparisons could be 435^9 (see figure 4). This research involved each judge ranking 15 packs of three scripts for each grade boundary, thus resulting in 45 paired comparison per judge. With six judges, this created 270 paired comparisons for each grade boundary, of which 210 were unique.

The resulting pack design template was used for each key grade boundary set of packs.

This resulted in a pack design with the following attributes:

⁹ The number of pairs generated from n objects is $n(n-1)/2$

Table 1: summary of pack design.

Packs per judge per key grade boundary	15
Number of scripts per pack	3
Number of comparisons per pack	3
Number of scripts per judge per key grade boundary	45
Number of comparisons generated per judge per key grade boundary	45
Total number of comparisons per key grade boundary	270
Number of different scripts per judge per key grade boundary	30
Percentage of comparisons <i>within</i> session	14%
Frequency of exposure to same script per judge	<i>Min</i> 1 <i>Max</i> 2
Frequency of same script exposure overall	<i>Min</i> 8 <i>Max</i> 10
Total number of possible unique comparisons	435

script n		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
year		4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6	7	7	7	7	7	
session		Jun	Jun	Jun	Jun	Jun	Jan	Jan	Jan	Jan	Jan	Jan	Jun	Jun	Jun	Jun	Jan	Jan	Jan	Jan	Jan	Jun	Jun	Jun	Jun	Jun	Jan	Jan	Jan	Jan	Jan
1	2004 June	o	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
2	2004 June		o	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57
3	2004 June			o	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84
4	2004 June				o	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110
5	2004 June					o	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135
6	2005 January						o	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159
7	2005 January							o	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182
8	2005 January								o	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204
9	2005 January									o	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225
10	2005 January										o	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245
11	2005 June											o	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264
12	2005 June												o	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282
13	2005 June													o	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299
14	2005 June														o	300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315
15	2005 June															o	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330
16	2006 January																o	331	332	333	334	335	336	337	338	339	340	341	342	343	344
17	2006 January																	o	345	346	347	348	349	350	351	352	353	354	355	356	357
18	2006 January																		o	358	359	360	361	362	363	364	365	366	367	368	369
19	2006 January																			o	370	371	372	373	374	375	376	377	378	379	380
20	2006 January																				o	381	382	383	384	385	386	387	388	389	390
21	2006 June																					o	391	392	393	394	395	396	397	398	399
22	2006 June																						o	400	401	402	403	404	405	406	407
23	2006 June																							o	408	409	410	411	412	413	414
24	2006 June																								o	415	416	417	418	419	420
25	2006 June																									o	421	422	423	424	425
26	2007 January																										o	426	427	428	429
27	2007 January																											o	430	431	432
28	2007 January																												o	433	434
29	2007 January																													o	435
30	2007 January																														o

Figure 4: Matrix displaying the 435 possible numbered comparisons between the 30 archive scripts on any key grade boundary. Blue type indicates comparisons where scripts are from the same session.

2.5 Procedure

Each judge was sent all materials by post in order that they could conduct the rank order task at home¹⁰. Materials consisted of:

30 packs of 3 scripts each: 15 grade A packs and 15 grade E packs

- A record sheet for each pack
- instructions
- Question papers for each session of Psychology Unit 2540 (not required for Economics 2882 which has a combined question and answer booklet).

Judges were instructed to rank the three scripts in each pack (see instructions in Appendix A). They were informed that one set of 15 packs contained scripts on or around the A boundary, and that the other set contained scripts on or around the E boundary, and that these were drawn from the last six sessions. For the purposes of control (i.e. to cancel out any possible confounding variables that might result from order effects), in both Psychology and Economics exercises, half of the judges were instructed to rank the E grade packs first, while half were instructed to rank the A grade packs first. Each script was labelled with the session, as well as a script code e.g. "January 2006 P013".

The judges were asked to place the scripts in rank order, from best to worst, based on a holistic judgement of the quality of the candidates' answers, and that any method might be used to do this except re-marking.

Judges were informed that each pack would contain scripts "reasonably close to one another when marked conventionally", but that the order of the packs, and the order of scripts within the packs were random. They were discouraged from making tied-rankings, but that, should they feel two scripts were genuinely of the same quality, that they could indicate this by drawing a bracket around them on the recording form.

They were given several weeks to complete the task to allow for flexibility. However, it was suggested that the task would take the equivalent of about one working day.

As a further task, to be carried out after ranking their 15 packs, judges were also asked to rank the question papers in order of difficulty (see Appendix B). The results for this part of the exercise are reported in section 3.2.

2.6 Analysis

The four sets of data (Psychology 2540 E grade, Psychology 2540 A grade, Economics 2882 E grade and Economics 2882 A grade) were each analysed separately by fitting a Rasch model (Andrich, 1978). The model fitted in this case was:

$$\ln[P_{ij} / (1-P_{ij})] = B_i - B_j$$

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

¹⁰ Black and Bramley (2008 in press) showed that a postal replication of a study conducted in a meeting provided very similar if not identical outcomes. Therefore, in order to avoid problems of trying to schedule a mutually convenient day, minimise cost and maximise number of judges and units, sending out materials by post was considered viable.

This analysis produces a measure for each script on a latent trait scale. The unit of the scale created by the analysis is known as a 'logit' or 'log-odds unit'. The analysis was carried out with FACETS software (Linacre, 2005).

3. Results

3.1 Analysis of Rank Order outcomes

The estimated measures and fit statistics for scripts and judges were considered. A detailed explanation of these statistics is not the focus of this report, but two points are worth noting:

- All four analyses created a meaningful scale with separation reliabilities (analogous to Cronbach's Alpha) above 0.75 – in other words the differences between the script measures were not solely attributable to measurement error (as would have been the case if the judges had, for example, tossed a coin to make their rankings).
- Two scripts at the Economics E grade boundary (one from June 2004 and one from June 2005) were eliminated from the analysis because they were always ranked lowest in the packs in which they appeared. The estimated measures for these scripts are extrapolations based on the other measures.

The mean measures of script quality for the five scripts from each examination session are plotted below in Figure 5 for each of the four key grade boundaries (i.e. Psychology grades A and E and Economics grades A and E).

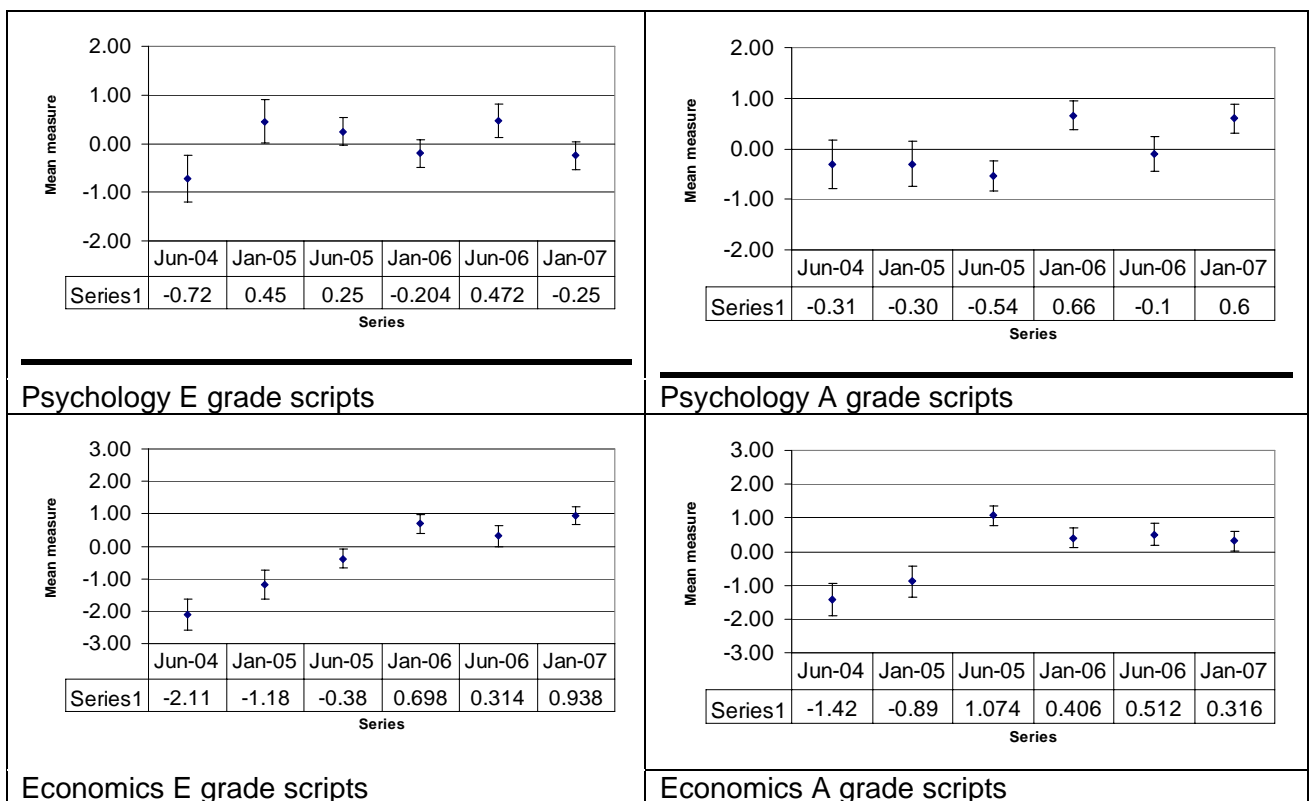


Figure 5: Means and measurement error bars plotted for each session¹¹.

¹¹ Please note that the scales on each of the four plots are not comparable. Here, a measure of zero denotes the notional mean measure of script quality for each of the four plots.

The 'error bars' in Figure 5 do not represent the usual standard error of the mean based on sampling variability, nor the within-session standard deviation, but represent $\pm 2 \times$ the measurement error in the mean of these particular scripts (i.e. treating each set of five scripts as the whole population of interest). The measurement error is calculated as:

$$\text{measurement error} = \frac{\sqrt{\sum_i se_i^2}}{n}$$

where se_i is the standard error of the estimated measure of script i .

This way of presenting this kind of data was suggested by Bramley (2007). The measurement error bars represent the uncertainty around each mean measure. Where there are overlapping values between two error bars, this represents the possibility that overall quality of the archives was similar (and that any apparent difference in the mean was due to measurement error).

Even accepting that there will be some variability in script quality on any particular mark, notionally, the *average* script quality on a key grade boundary should theoretically be the constant session upon session. In this exercise, a plot depicting a constant mean measure for each session would indicate a perfectly identical standard. The charts in Figure 5 do seem to indicate that, even once measurement error is accounted for, there are differences in the overall perceived standards of archives between sessions. Interestingly, for each of the four boundaries under consideration, all show a net increase in the quality of scripts, suggestive of an increase in standards over time. This is most consistently depicted for Economics, where mean script quality does appear to have risen, both for E and A grade overall. This view was also expressed by the Chief Examiner for this specification (prior to this data being keyed or analysed). He attributed this rise in performance standards to an increase in specification specific resources, such as text books, as well as effective INSET programmes for teachers, resulting in greater familiarity with the demands of the assessments.

An analysis of variance (ANOVA) can be used to compare the means of different groups taking into account within-group variability for each of the four grade boundaries in the study. These tests treat the scripts as random samples from some defined population, and the estimated script measures as known point values with no measurement error.

A one-way ANOVA for this data reveals there to be a significant difference in script quality across the different sessions for both Economics E grade scripts ($p=0.011$) and Economics A grade scripts ($p=0.001$). In Psychology, in contrast, the hypothesis that all the sessions had the same mean measure was not be rejected at the 5% level (i.e. $p>0.05$).

In order to analyse whether January and June sessions are awarding independently, it is necessary to conduct a two-way analysis of variance which includes an interaction term between session type and order.

Figure 6 shows plots of marginal means for each session estimated from the model stated above, where the blue line plots, in sequence, each of the three June sessions and the green line plots each of the three January sessions, therefore enabling comparison between January and June.

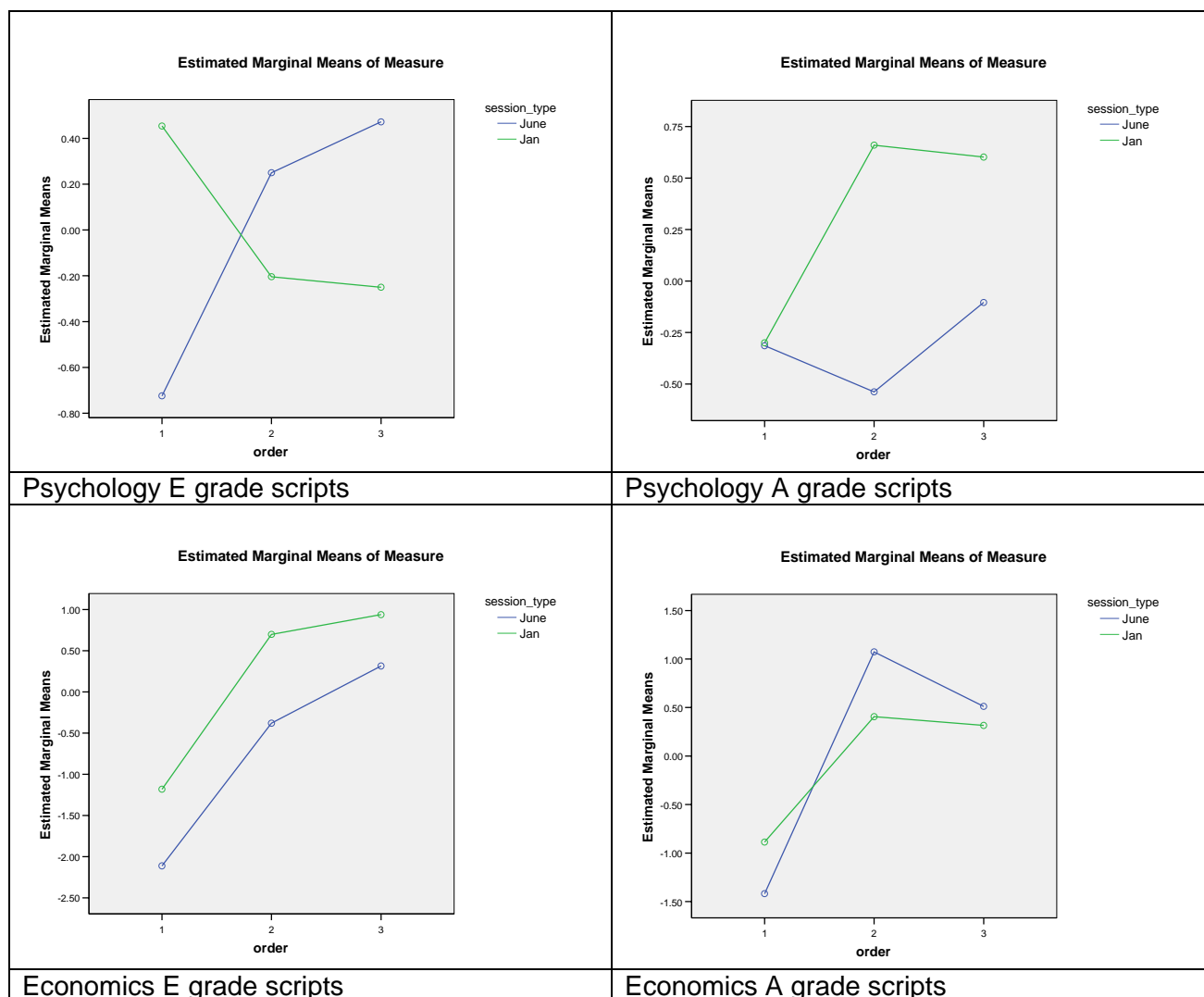


Figure 6: Plot of marginal means of script quality by order of session, comparing January and June sessions.

An interesting contrast is revealed between the Psychology and Economics scripts. The Economics scripts essentially depict the January sessions as shadowing or tracking the June sessions, following the same trend. (N.B. the effect of the order of the sessions is significant for both E and A scripts, ($p=0.01$ and $p=0.011$ respectively in line with the results of the one-way ANOVA).

The plots for Psychology contrast with those of Economics. While these plots suggest an interaction effect between session type (January versus June) and the order of the sessions, it must be noted that this difference was non-significant. ($p=0.253$ and $p=0.586$) (Full tables of ANOVA output for the one-way and two-way analyses are in appendices F* and G* respectively).

In terms of the original hypothesis, that January and June awarding sessions operate independently, though it might appear these charts might provide some support to this, it must be noted that the findings are non-significant.

3.2 Ranking question paper difficulty

Another aspect of this research deals with the issue of question paper demand. Judges were asked, after completing the script ranking exercise, to rank the question papers in order

of difficulty. This is interesting in that judges may find it difficult to take into consideration question paper difficulty when judging script performance.

Does this, in any way explain apparent differences in judgments about script standards?

Table 2a: Mean rankings of question paper difficulty for Psychology, (a higher score indicates greater perceived question paper difficulty).

Psychology 2540 question paper difficulty									
Overall rank	Session	Mean judged rank	s.d.	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6
1	June 2004	1.00	0.00	1	1	1	1	1	1
2	Jan 2006	2.83	0.98	2	4	2	2	4	3
3	June 2005	3.50	1.22	5	3	3	5	3	2
4	Jan 2005	3.83	0.98	3	3	4	3	5	5
5	Jan 2007	4.83	1.60	6	6	5	6	2	4
6	June 2005	5.17	0.98	4	5	6	4	6	6

Table 2b: Mean rankings of question paper difficulty for Economics.

Economics 2882 question paper difficulty									
Overall rank	Session	Mean judged rank	s.d.	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6
1	June 2006	2.08	1.74	1	1	1	5	1	3.5
2	June 2005	2.17	1.47	2	2	5	1	2	1
3	June 2004	3.67	1.97	6	6	2	2	4	2
4	Jan 2006	4.08	1.20	5	4	6	3	3	3.5
5	Jan 2007	4.33	1.37	4	3	3	4	6	6
6	Jan 2005	4.67	1.03	3	5	4	6	5	5

For Psychology, the June 2004 was unanimously the easiest paper (see Table 2a above). There was generally more agreement amongst the Psychology judges than the Economics judges, as seen in judge rankings (and represented by the standard deviation, see Tables 2a and 2b) possibly indicating that, for Economics, the differences in question paper difficulty were small and therefore difficult to rank. (Indeed, one Economics judge commented that this was the most difficult task of the whole research).

Is there any notable relationship between judgements on question paper difficulty and performance on papers? One might conjecture that judges, though instructed to compensate for any differences in question paper difficulty when making judgements about script quality, might not be able to make adequate compensation. Some researchers have expressed such doubts: *"It is, after all, the central conundrum of testing educational attainment: the relative merits of easier tasks done well and harder tasks done moderately."* (Adams & Pinot de Moira, 2000). If it were the case that judges could not make adequate compensation, one might posit that scripts based on harder questions papers might receive lower measures than scripts based on easier question papers and that a correlation would be produced.

Table 3; Table showing correlations (Spearman's Rho) between mean script measures and mean question paper difficulty ranks:

Scripts	Correlation coefficient (Rho)	significance

Psychology E scripts	0.31	0.54
Psychology A scripts	-0.26	0.62
Economics E scripts	-0.09	0.87
Economics A scripts	-0.60	0.21

However, this is not the apparent in the correlations between mean question paper difficulty ranking and the mean measure produced for each session (see Table 3). This seems to indicate that there is no discernable relationship between question paper difficulty and judged script quality.

3.3 Judge feedback on the activity.

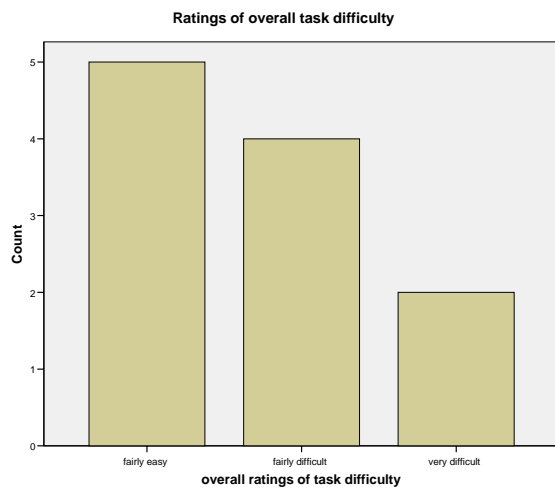
Judges were asked to complete a feedback form as soon as possible after completion of the main task, and return it either postally or electronically. The feedback form is given in Appendix C.

3.3.1 *Difficulty of the overall task*

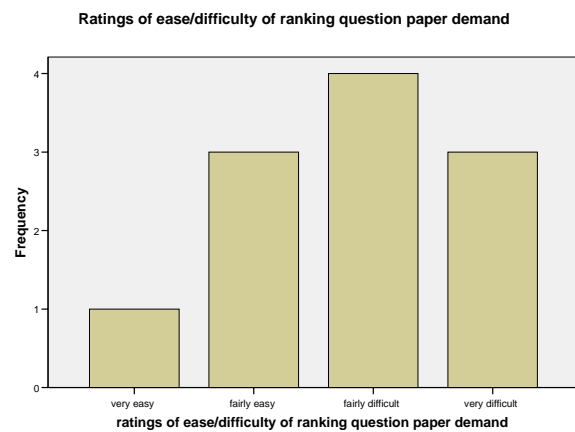
On a four point scale (1 = very easy; 4 = very difficult), they were asked to rate the difficulty overall of the task. The modal responses was “fairly easy”, though 6/11 respondents found the task difficult to a degree. There was no notable difference between Psychology and Economics judges. Some commented upon how the task became easier over time. Sources of task difficulty recorded included the similarity of script quality (n=2) and the difficulty of making comparison between scripts from several sessions (n=3). According to the judges, the mean time taken per pack was 12 minutes, though the Psychology judges notably faster than Economics judges (9 minutes compared to 14 minutes). This might be explained by the Psychology judges being more familiar with this type of exercise. A number of judges remarked that their speed increased with task familiarity.

3.3.2 *Allowing for differences in question paper difficulty*

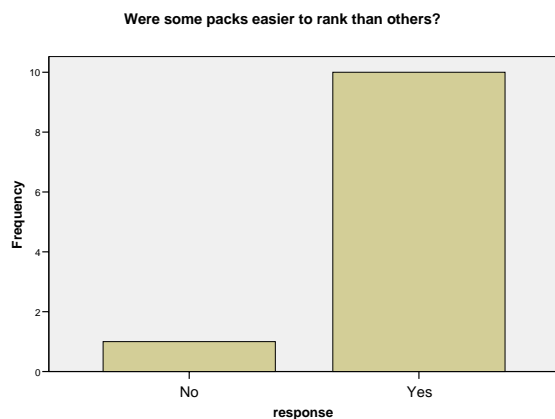
This is the first exercise we have conducted where judges have been asked to rank order scripts from more than two sessions. The judges were asked “how easy or difficult did you find taking into account differences in question paper difficulty”. The modal response (see Figure 7 below) was “fairly difficult”. One (psychology) judge responded: “It creates so many variables! Also, I had to keep reminding myself what they should actually be writing about... i.e. answers I thought looked right were wrong when I checked”. Another examiner noted: “The differences between sessions seem minimal to me. It is swings and roundabouts since some questions are slightly easier but others slightly harder and overall it averages out much the same.” Another suggested that making compensations for question paper were not particularly relevant: “... Quality and extent of answer can be done (= “*judged*”?) without (looking at the) question”. In contrast, one judge commented: “...(it) did not always seem to be comparing like with like. Is a good brandy better than a good whisky?”. Interestingly, one might predict that the Economics judges may have found this aspect of the task slightly easier than the psychology judges due to the combined question and answer booklet (negating the need to refer to multiple bits of paper). However, this was not the case.



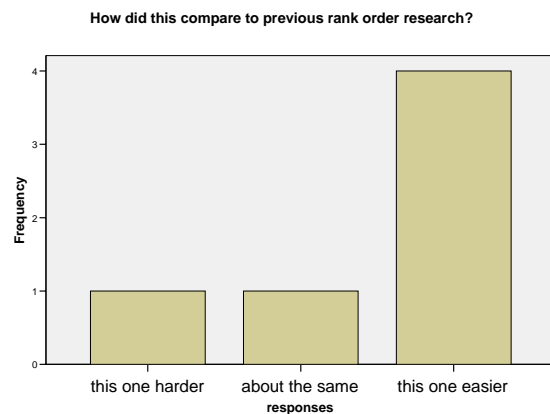
a) Ratings of overall task difficulty (n=11)



b) Ratings of ease/difficulty of ranking question paper demand (n=11)



c) Responses to whether some packs were easier to rank than others (n=11)



d) Comparison of ease /difficulty of *this* rank order exercise compared with previous rank order exercises (involving 10 scripts per pack, wider raw mark range) (n=6)

Figure 7: Bar charts of judge feedback.

3.3.3 Were some scripts easier to rank than others?

The vast majority of judges responded that some packs were easier to rank than other packs (see figure 7c). Judges were asked to explain why some packs were easier than other.

Factors which made ranking *easier* included:

- Packs containing two of the three scripts from the same session (n=4)
- Where there were more discernable differences in script quality (n=3)
- E grade packs (n=2)
- A grade packs (n=2)

Conversely, factors which made the rank-ordering more difficult included:

- Perceived closeness of standard (n=3)
- Packs containing three scripts were from three different sessions (n=2)
- Comparing scripts exhibiting inconsistent performance (n=2) (Consistent with Baird, 2000)
- E grade packs (n=1)
- A grade packs (n=2)

3.3.4 Rank ordering strategies

Judges, in their instructions (see Appendix A) were encouraged to determine their own strategy. They were asked to describe their adopted strategy. Some strategies (or elements of strategies) were:

- Focusing on 'key', discriminating questions (n=5)
- Skimming whole scripts (n=4)
- A provisional ranking preceding a more definitive ranking. (n=4)
- General language and expression (n=2)
- Terminology (n=4)
- Correctness / incorrectness of responses (n=5)
- Consistency versus inconsistency of responses (n=4)
- Expansion of points/length of responses (n=4)
- Evidence of understanding rather than just knowledge (n=3)
- "quality of answer" (n=1)
- "knowledge of subject" (n=2)
- relevance (n=1)
- "technical quality" (n=1)
- quality of evaluation (n=4)
- quality of diagrams (relevant only to Economics) (n=3)
- ability to analyse and structure arguments (n=1)
- discursive qualities (n=1)
- using differing criteria for A scripts versus E scripts (n=2)

Many examiners' strategies involved more than one element, or, multiple heuristics. For example:

I first identified 'key' questions on each paper that I thought would discriminate well... when I looked at the scripts, I compared responses on the 'key' questions and made a first ranking, bearing in mind the respective difficulty of those questions. I then checked the ranking in order by skimming the whole script.

In some it is clear that judges employ, as above, a refining process, and/or compensatory models, where one element of performance may be compensated by another e.g:

I began by looking at the whole script to see if all questions were answered. Gaps would mean I would initially place that script bottom. Then, I looked at one page of each script question by question. I always mentally credited understanding over knowledge. And if they(the scripts) were equal I then looked for terminology. At this stage, it did not matter which questions I was comparing it was just an impression of the three pages side by side. If I could not separate them on that basis, I then turned to the next page and looked at specific questions... which seemed the best discriminators

Some strategies may be more effective, more valid, or simply easier to deploy. This is one possible area for further research. To date, judges have, by instruction, determined their own strategies. The effects of intervention, instruction or training could be investigated. It is worth bearing in mind that information on the strategies used by judges in this rank-ordering exercise is naturally limited by the data collection method (i.e. post hoc self-report). A more sophisticated data collection technique appropriate here would be verbal protocols, as this allows 'live' or concurrent reporting by the judges of their activities in fuller detail.

Whilst this may cause some interference with conducting the actual rank-order task, it is not as prone to certain deficiencies of post-hoc self-report such as self-serving memory biases.

3.3.5 Differences and similarities between making judgements in the context of rank-ordering versus awarding.

Judges were asked to comment upon the way in which their judgement of an individual script's quality was different in this exercise compared to during an award. According to the judges, judgements undertaken in an awarding context are different in a number of ways:

Table 4: differences between making script judgements in awarding versus rank ordering, according to judges.

Awarding	Rank ordering	n=
Key discriminating questions are identified by the PE.	Key discriminating questions are identified by the judge.	4
Scripts are compared to 'internalised' standard from the archive.		1
Scripts are compared to published performance criteria (e.g. QCA)		1
Awarding involves absolute judgements (e.g. grade A or not grade A).	Rank-ordering involves relative judgements, comparing one script with another.	1
Might be looking at individual questions more than whole scripts	More holistic judgement	1
	More time available or given to each script in rank-ordering	3
Having scripts with annotations diverts attention to most relevant parts of an answer – therefore easier than rank-order		1
Only one mark scheme / question paper at a time – easier.		1
Dialogue between awarding committee members may have an influence		1

Other comments:

- For rank ordering. can open up all the scripts at the same time at home [though this would be true of remote awarding also] n=1
- Awarding and rank-ordering (in this exercise) were similar (“Is it an A or not?”; “very little difference”; “I look at key questions in both”) (n=3)
- “I do believe that both of these are important in the awarding process and lead to appropriately determined grade boundaries”.

Finally, the psychology judges were asked about how the relative task difficulty of this rank-order exercise compared with previous (involving ranking 10 scripts per pack, from 2 sessions, with a wider raw mark range). Of the six judges, five found this rank-order exercise easier. Usually the explanations provided indicated that holding the standard of 10 scripts in one's mind was much more difficult, even acknowledging that, in this exercise, the scripts were much closer in standard.

4. Discussion

4.1 Limitations of this research

There are several limitations of this study which should be mentioned. First, only two units were analysed. Thus, the generalisability of this research to other units and other specifications is unclear. Given that the two units in question show differing patterns, it is unknown what patterns other units might display.

Second, having only three sessions of each of January and June is only barely sufficient to show the beginnings of a trend. It would be preferable to have perhaps four or more sessions.

Third, whilst the archives are chosen in order to be representative of the standard at the key grade, there are only five scripts on the mark point. Access to a broader archive, either more scripts on the mark point and/or a greater range of scripts on different mark points might allow for more secure outcomes.

Fourth, whilst increasing performance standards may have been detected in this research over time, it is not possible to conclude what impact, if any, this had upon specification level outcomes. Indeed, when awarding units, there is a tension between unit level outcomes and specification level outcomes. It is possible that the awarding committee may utilise some sort of compensatory model, whereby if the outcomes are slightly high in one unit, this may be offset by a more stringent decision in another unit.

Finally, the task asked of judges in this exercise could be viewed as difficult and demanding. They were asked to rank scripts that were notionally very close in standard, whilst taking account of six different question papers with likely differences in overall assessment demand. Previous Thurstone and rank order activities tend to have involved comparisons between only two question papers or specifications at any point and to have ranked scripts with a greater range of raw mark. Feedback from the judges, as described above, is relevant here. However, in terms of the minimal number of misfitting judgements, any difficulty encountered has not produced aberrant rankings.

4.2 Other issues

Comparability research often raises issues to do with archive scripts. Archives are the really the 'Long Term Memory' of standards performance. But perhaps this memory is too selective (only five key grade boundary scripts). An archive of one or two scripts on every mark point would capture a fuller record of an assessment and increase the possibilities of comparability research for the future. This would also facilitate research which could quantify differences in standards in terms of raw marks. However, there would be storage and cost implications of such an archive.

It would be possible to take a different approach to the data analysis, and instead of separating the measurement aspect (the construction of a scale via the paired comparisons) from the statistical analysis (the ANOVAs), to combine these within a single general modelling approach – for example using logistic regression, as advocated by John Bell in *Greatorex et al.* (2002). The pros and cons of this are discussed in Bramley (2007).

Further research could be undertaken to look at January versus June awarding standards, either upon individual units or indeed whole specifications. This may further determine the extent and nature of January versus June awarding standards. This might even be considered essential if awarding protocols for January are in need of review.

As rank order is becoming an increasingly useful method for investigating comparability, further investigation into the effectiveness of different rank-ordering strategies should be undertaken.

In terms of the use of Thurstone Triples (as first used in this research study), other assessment research and operations are already adopting this technique. Firstly, Nick Raikes et al (2008) have in recent research used Thurstone Triples with considerable success as a method by which to extend the pool of judges involved in awarding judgements for a biology AS unit. Secondly, Professor Richard Kimbell's Project E-scape project will be trialling triples as part of the way in which GCSE Design and Technology e-portfolios will be assessed.

References

- Adams, R. & Pinot de Moira, A. (2000). *A Comparability Study in GCSE French including parts of the Scottish Standard Grade Examination. A study based on the Summer 1999 examination*. Organised by WJEC and AQA on behalf of the Joint Forum for GCSE and GCE.
- Baird, J-A. (2000). Are examination standards all in the head? Experiments with examiners' judgments of standards in A level examinations. *Research in Education*
- Baird, J-A. (2007). Alternative conceptions of comparability. In P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.
- Black, B., & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education* 23 (3).
- Bramley, T. (2005). A Rank-Ordering Method for Equating Tests by Expert Judgement. *Journal of Applied Measurement*, 6 (2), 202-223.
- Bramley, T. (2007). Paired comparison methods. In P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.
- Bramley, T., & Black, B. (2008). *Maintaining performance standards: aligning raw score scales on different tests via a latent trait created by rank-ordering examinees' work*. Paper presented at the Third International Rasch Measurement conference, University of Western Australia, Perth, January 2008.
- Gill, T., Bramley, T., & Black, B. (in prep.). An investigation of standard maintaining in GCSE English using a rank-ordering method.
- Greatorex, J., Elliott, G. & Bell, J.F. (2002). *A Comparability Study in GCE AS Chemistry*. A study based on the Summer 2001 examination and organised by the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for OCR on behalf of the Joint Council for General Qualifications.
- Linacre, J.M. (2005). Facets Rasch measurement computer program. Chicago: Winsteps.com.

Newton, P.E. (1997). Measuring comparability of standards between subjects: why our statistical techniques do not make the grade. *British Educational Research Journal*, 23(4), 433-450.

QCA (2008). GCSE, GCE, VCE, GNVQ and AEA Code of Practice, London, QCA.

Raikes, N., Scorey, S. and Shiell, H. (2008) Grading examinations using expert judgements from a diverse pool of judges. A paper presented to the 34th annual conference of the International Association for Educational Assessment, Cambridge, UK.

Schagen, I., & Hutchison, D. (2007). Multilevel modelling methods. In P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Appendix A: Rank Ordering Instructions

Instructions for completing the rank-ordering task on Economics 2882. *Please read carefully.*

You have thirty separate packs of three 2882 scripts. Packs 1 to 15 of your packs contain scripts on or around the A grade boundary. Packs 16 to 30 contain scripts on or around the E grade boundary.

The scripts have been drawn from the last six sessions (June 2004 to January 2007 inclusive). The scripts you receive in each pack are in no particular order, and have been cleaned of marks.

The task described below should be carried out thirty times; once for each pack of scripts. You do not have to complete all thirty packs in the same sitting! You will probably want to break the task up into more manageable chunks!

The task we would like you to complete is to place the 3 scripts in each pack into a single rank order from best to worst, based on the quality of the candidates' answers. You may use any method you wish to do this, based on scanning the scripts and using your own judgement to summarise their relative merits, but you must not re-mark the scripts. You should endeavour to make a holistic judgement about each script's quality. Remember, this is not a re-marking exercise.

No tied ranks are allowed. If you are concerned that two or more scripts are genuinely of exactly the same standard you may indicate this by placing a bracket around them on the record sheet, but you must enter every script onto a separate line of the record sheet.

Whilst it can be difficult to make relative judgements about scripts from different sessions, we ask that you do this as best you can, using your own professional judgement to allow for differences in the questions and stimulus material. We have provided a copy of each of the question papers in order to assist in this. We suggest that before you begin the task, you re-familiarise yourself with the question papers from all the years.

Each pack contains scripts which are reasonably close to one another when marked conventionally. The order of the packs is arbitrary. Similarly, you should make no assumptions about the way in which the scripts from the different years are ordered within each pack. They are deliberately randomised.

Once you have decided upon a single rank order for the three scripts from a pack, please record the order on the sheet enclosed in the pack using the script I.D. (e.g. Ec026), and return the scripts to the pack before beginning another 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.

Finally, when you have completed ranking all 30 packs, there are two final, short tasks. Firstly, complete the task in the sealed envelope; secondly please complete the feedback questionnaire emailed to you/posted to you separately late March. Thank you!!

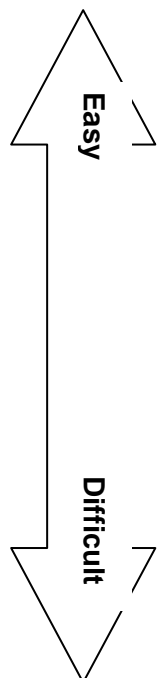
Appendix B: Question Paper ranking activity

You must have completed rank ordering scripts in the 30 packs before commencing this task.

Now consider the relative difficulty of each *question paper* as a whole. On balance, you may think that some question papers are more easy or more difficult than others.

Please place these question papers in rank order, from (rank 1) most easy/accessible to (rank 6) most difficult/inaccessible. If you think that two question papers are genuinely of the same difficulty, you may indicate this by placing a bracket around them on the record sheet, but you must enter every script onto a separate line.

Please record this in the table below.



Question Paper		
	Session (January/June)	Year
1.		
2.		
3.		
4.		
5.		
6.		

Appendix C: Example Feedback Questionnaire

Feedback Questionnaire for Rank Order 2540

We are very interested in your experiences of this task.

Please could you answer the following questions:

1. How easy or difficult overall was the task?

<input type="checkbox"/>	very easy
<input type="checkbox"/>	fairly easy
<input type="checkbox"/>	fairly difficult
<input type="checkbox"/>	very difficult

Please expand your answer:

2. On average, about how long did it take to rank order the scripts in each pack?

_____ minutes per pack.

3. The task involved comparing scripts from different sessions. How easy or difficult did you find taking into account differences in question paper difficulty?

<input type="checkbox"/>	very easy
<input type="checkbox"/>	fairly easy
<input type="checkbox"/>	fairly difficult
<input type="checkbox"/>	very difficult

Please expand your answer:

4. Did you find some packs of scripts were more easy / difficult to rank than other packs?

<input type="checkbox"/>	No – the packs were all about the same ease/difficulty to rank.
<input type="checkbox"/>	Yes – some packs were easier/more difficult than others.

Please go to question 5.

Please go to question 4b.

- 4b. Please briefly describe or identify the packs which contained scripts which were **easier** to rank **and** explain why you found them easier to rank.

4c. Please briefly describe or identify the packs which contained scripts which were **more difficult** to rank **and** explain why you found them more difficult to rank.

5. Please can you describe the process or strategy you used for this task, for example: how did you decide that one script was better or worse than another? What were the features of the scripts which influenced your judgements?

6. How does the way you make judgements about an individual script's quality in this exercise differ from the way you make judgements about scripts during an Awarding Meeting?

7. How did this task compare to the previous rank order study which involved ranking 10 scripts per pack? Tick the box which best represents your view:

This rank order task (3 script packs) **was harder** to complete than **the previous rank order task(s)** (10 script packs)

This rank order task was about **as easy/difficult** to complete as **the previous task(s)**

This rank order task was **easier** to complete than **the previous task(s)**.

Please briefly explain the reasons for your view:

Once again, many thanks for participating in this study.
Please return this questionnaire along with the other materials.

Appendix D Tables of One-way ANOVA results.

Psychology E grade scripts

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Measure * series	Between Groups	(Combined)	5.598	5	1.120	.616	.688
	Within Groups		43.599	24	1.817		
	Total		49.198	29			

Psychology A grade scripts

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Measure * series	Between Groups	(Combined)	6.434	5	1.287	.796	.563
	Within Groups		38.785	24	1.616		
	Total		45.220	29			

Economics E grade scripts

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Measure * series	Between Groups	(Combined)	34.862	5	6.972	2.654	.048
	Within Groups		63.061	24	2.628		
	Total		97.923	29			

Economics A grade scripts

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
Measure * series	Between Groups	(Combined)	22.380	5	4.476	3.839	.011
	Within Groups		27.985	24	1.166		
	Total		50.365	29			

Appendix E Tables of Two-way ANOVA results

Psychology E grade Scripts

Tests of Between-Subjects Effects

Dependent Variable: Measure

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	5.598(a)	5	1.120	.616	.688
Intercept	3.33E-006	1	3.33E-006	.000	.999
session_type	3.33E-006	1	3.33E-006	.000	.999
order	.311	2	.155	.086	.918
session_type * order	5.288	2	2.644	1.455	.253
Error	43.599	24	1.817		
Total	49.198	30			
Corrected Total	49.198	29			

a R Squared = .114 (Adjusted R Squared = -.071)

Psychology A grade Scripts

Tests of Between-Subjects Effects

Dependent Variable: Measure

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	6.434(a)	5	1.287	.796	.563
Intercept	3.00E-005	1	3.00E-005	.000	.997
session_type	3.066	1	3.066	1.897	.181
order	1.600	2	.800	.495	.616
session_type * order	1.769	2	.884	.547	.586
Error	38.785	24	1.616		
Total	45.220	30			
Corrected Total	45.220	29			

a R Squared = .142 (Adjusted R Squared = -.036)

Economics E grade Scripts

Tests of Between-Subjects Effects

Dependent Variable: Measure

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	34.862(a)	5	6.972	2.654	.048
Intercept	2.477	1	2.477	.943	.341
session_type	5.773	1	5.773	2.197	.151
order	28.821	2	14.410	5.484	.011
session_type * order	.268	2	.134	.051	.950
Error	63.061	24	2.628		
Total	100.400	30			
Corrected Total	97.923	29			

a R Squared = .356 (Adjusted R Squared = .222)

Economics A grade Scripts

Tests of Between-Subjects Effects

Dependent Variable: Measure

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	22.380(a)	5	4.476	3.839	.011
Intercept	1.33E-005	1	1.33E-005	.000	.997
session_type	.092	1	.092	.079	.781
order	20.461	2	10.230	8.774	.001
session_type * order	1.827	2	.914	.784	.468
Error	27.985	24	1.166		
Total	50.365	30			
Corrected Total	50.365	29			

a R Squared = .444 (Adjusted R Squared = .329)