FHS Physics 2006

Chairman's Report

Paper Preparation The number and format of the papers was unchanged from last year. In an attempt to achieve more expert and more thorough checking of papers, checkers were used in series rather than in parallel, and were required to work the papers in full without access to the model solutions. When significant work was done on a paper in response to a checker, a new checker was engaged for the next check. These procedures were extended to Short Option papers. To offset the extra labour inherent in these procedures, there were no plenary paper readings. Overall the new system saves some Examiner time, and probably improves paper quality.

The problem of circulating drafts without compromising security is a significant issue – some minor errors would have been picked up if it had been easy to show a near final draft to the right person. It is not clear whether this problem can be solved electronically; the prospect of the papers appearing on the Web prematurely is so ghastly that we should probably make more systematic use of secretarial delivery.

Table 1: Statistics of individual papers. All papers out of 100 except S papers, which are out of 50

	A1	A2	A3	B1	B2	B3	C1	C2	C3	C4	C5	C6
Candidates	164	164	164	147	138	147	54	38	46	43	21	42
Target mean	65	65	65	65	65	65	63.56	64.97	62.60	65.93	61.61	70.19
Mean	53.63	55.22	59.59	66.58	56.02	55.22	58.58	62.74	60.96	69.49	65.00	60.58
SD	11.75	15.62	15.06	16.55	15.81	16.21	14.05	17.60	15.50	17.40	11.80	14.67

	S1	S2	S4	S6	S7	S8	S9	S10	S12	S15	S16	S17	S18	S22
Candidates	29	46	95	8	19	11	41	26	25	17	12	13	5	10
Mean	36.03	32.17	36.10	33.31	36.42	39.18	33.71	24.39	28.76	32.71	38.58	33.46	30.60	35
SD	12.30	8.71	6.96	2.49	14.55	8.22	9.46	9.26	7.85	8.42	7.33	7.53	11.15	5.66

Note: S22 is the option Teaching and Learning Physics in Schools

Table 1 shows the means and standard deviations of individual papers. The data for the A and B papers show that we are having difficulty adjusting to the increase from 55 to 65 in the target mean mark: with the exception of B1, these papers come close to the old target. The "target mean" for the C papers differs from 65 for reasons explained below under the heading "scaling". These papers come close to target, the exception being C6, which has a very high target mean because this paper attracts such strong candidates.

The Examinations The Examinations went off without major incident, in large measure because the Examination Schools constitutes a notably efficient machine.

The number of candidates who do not sit in the main hall, because they have extra time or special requirements, seems to increase continually. These candidates create significant extra work for Examiners because (a) their scripts are not available in Schools for up to 24 hours after the examination, and (b) their scripts are then filed in a confusing way: scripts from the same paper will be filed under different headings depending on the candidate's year and FHS. For example, scripts for Short Option S1 might be filed under Prelims, Part A, Part B BA, Part B MPhys, Part A P&P or Part B P&P. Similarly, a script for B1 might be filed under Part B BA, Part B MPhys or Part B P&P. Each of these possibilities is known to Schools only as an instantly forgettable code, such as XPHC, DPHD, DPHE, CPSC, etc. *Much would be gained, and probably nothing lost, if Schools treated all scripts for a given paper in the same way.*

On account of staffing problems in Schools, we received the first candidate lists only very shortly before the examination started. In particular, the marklists for the Short Options were not ready when the I went to Schools to split the ~ 500 scripts into bundles for the 12 examiners. Without lists of candidates taking each paper, it was impossible to check the bundles, and the upshot was that too many solutions went to the wrong examiner. This problem led to significant confusion and waste of Examiners' time.

Marking Marks for all 24 papers were returned by the deadline, which had been brought forward from last year by 24 hours. All scripts were centrally checked to ensure that (a) every page had been scanned by an Examiner, (b) that the marks noted in the margin were correctly summed to question totals on the cover

sheet, and (c) that these totals had been correctly entered on the spreadsheets. This checking process threw up a fair number of errors, overwhelmingly by one or two marks only.

Projects Candidates rose superbly to the challenge set by the new guidelines for the preparation of MPhys reports; the reports were much improved in standard and easier to read.

As last year, each essay or report was read by a randomly chosen examiner and a more expert Assessor, both grading on the same form, though revised forms were used. As hoped, the redesigned assessment forms narrowed the spread in the marks awarded: the Examiners' marks had mean of 72.7 and a SD of 9.93. An analysis of the marks assigned by Examiners (each of whom read 17–20 reports) suggested that there were systematic offsets in their marks. This hypothesis was tested by two Examiners from the middle of the distribution third-reading four reports originally read by one of the outlying Examiners. Since the results of these readings was consistent with the existence of offsets, it was resolved to scale each Examiner's marks to a common mean using the same quadratic scaling algorithm that was used to scale the A, B and C papers.

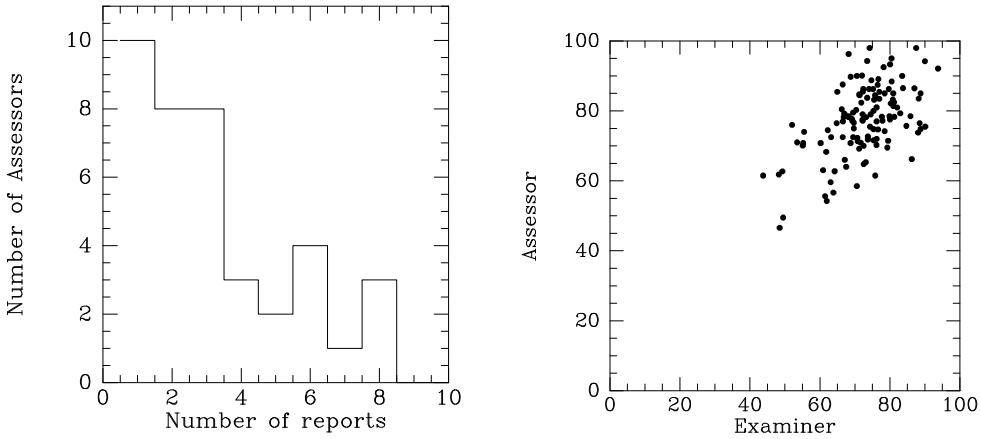

Figure 1: (a) number of reports read by Assessors; (b) scatter $A - E$ for MPhys projects

Although some Assessors read enough reports (> 6) to get some idea of the appropriate standard, far too many reports were read by Assessors who had read only one or no other reports (Figure 1a). A particularly worrying statistic was that the rms of $A - E$, where A and E are the Assessor and Examiner marks, respectively, is for some Assessors bigger than the rms of A. Figure 1 shows the degree of correlation of A and E.

In an endeavour to improve the quality of both A and E, Examiners contacted Assessors in cases when their marks differed by more than 15. While neither party was to feel pressure to change his/her mark, both parties were asked to reconsider their marks in the light of a discussion of what each liked and disliked about the report. Some but by no means all discussions lead to a convergence of marks.

It is impossible to feel confidence in Assessors' marks, and the Examiners were very glad that this year the Assessor's mark contributed 30% of the final mark, rather than 70% as last year. *Next year a greater effort must be made (a) to reduce the number of Assessors and (b) to clarify what they should be looking for in a report.*

While we must not deviate from the principle that we are not simply assessing the quality of the science in the report (because this is so heavily influenced by the nature of the project) this year's assessment form probably does not attach appropriate weight to the student's intellectual achievement in getting to grips with a difficult topic and learning challenging new physics.

BA Reports and Essays The Examiners were worried that assessments of BA reports would be adversely affected by comparison with MPhys reports. This comparison was invited in as much as the same form was used to grade both kinds of report. On the other hand, most Examiners read their BA reports first. *Next year a different questionnaire should be used for the two reports.*

The only new problem thrown up by the assessment of BA Essays was a case of plagiarism that was referred to the Proctors. The candidate was eventually penalized by the Summary Court of Jurisdiction.

Since each Examiner read only 4–6 BA reports or essays, scaling was not practicable. *Next year all BA reports should be read by one Examiner, who does not read any MPhys reports, and all BA essays should be read by another Examiner.*

Scaling As previously, marks from A, B and C papers were scaled. However, this year linear scaling to the University mandated mean of 65 was replaced by quadratic scaling that left 0 and 100 invariant. The target mean of a C paper was not 65 but 65*(B-paper mean of these candidates)/(B-paper mean of all MPhys candidates). The same algorithm was used to scale the MPhys project marks of individual examiners.

Table 2. Percentages in each class of the BA

Year	1	2.1	2.2	3	Pass	Fail
2006	23.7	23.7	26.3	18.4	7.9	0
2005	8.8	26.5	35.3	23.5	5.9	0
2004	20.5	46.2	25.6	7.7	0	0
2003	10	66,7	6.7	10	6.7	0
2002	12.5	40	30	12.5	5.0	0
2001	14.6	53.7	26.8	4.9	0	0
1996	9.8	28.8	21.6	7.8	2.0	0

Table 3. Percentages in each class of the MPhys

Year	1	2.1	2.2	3	Pass	Fail
2006	39.5	41.9	15.3	3.2	0	0
2005	35.7	47.6	11.9	4.8	0	0
2004	36.7	46.7	12.5	4.2	0	0
2003	34.9	53.2	6.3	5.6	0	0
2002	31.1	51.6	15.6	1.6	0	0
2001	35.8	53.3	8.3	2.5	0	0

Classification Candidates were classified as previously announced in the *Examination Conventions*. Overall percentages were calculated by dividing each candidate's weighted total mark by the maximum mark achievable by this candidate. Then candidates were ranked by decreasing value of this percentage, and lines were pencilled in just above 70%, 60%, 50%, 40% and 30%, avoiding locations where adjacent scores were extremely close. This procedure resulted in the class percentages given in Table 2 for the 37 BA candidates classified and in Table 3 for the 124 MPhys candidates classified.

Fluctuations in the statistics for the BA are expected to be significant, especially at the top end, which may include strong candidates who have decided to leave after three years for economic reasons. Table 2 shows that this year the proportion of BA candidates achieving a first was historically high. Table 3 shows the corresponding data for the MPhys. The first class may have grown slightly at the expense of the 2.1 class. The internal Examiners were worried by that this might represent grade inflation. The external examiners advised that at other institutions marks over 70% invariably earned a first, and the University's decision to increase the target mean to 65% was probably intended to increase the number of firsts awarded in line with practice elsewhere. At 81.4% the percentage achieving a "good degree" (1 or 2.1) is low historically.

Special cases All letters and certificates forwarded by the Proctors over the last three years were summarized on a spreadsheet that was circulated to all examiners in advance of the marks meeting. After class boundaries had been pencilled in, the Proctors' instructions on handling medical certificates were read, and then each candidate on the list was considered, and concensus achieved as to whether the candidate should change class. One candidate was permitted to migrate into a higher class, and another candidate was helped over the MPhys hurdle.

Data Handling This year data were processed by the system constructed by the Department's IT professionals rather than the Chairman's home-built system. As soon as I became Chairman I was aware that data handling was an issue, and was anxious to have a complete review of the system and a dummy run using the 2005 data before that start of HT. Unfortunately, my request for IT help was at first not taken

seriously, and the necessary coding was not done until the end of HT, by which time paper preparation and then Project and Essay grading left no time for the envisaged the dummy run.

As in recent years, Examiners returned marks on Exel spreadsheets. These devices do not inspire confidence.

- One problem arises when candidates that have withdrawn are left on the spreadsheet. Exel then assigns them a total of zero even when there are no entries in their row. These zeros then lead to the paper mean and standard deviation being incorrect. Withdrawn candidates need to be excised from marksheets.
- Errors arose because the formula associated with some cell was incorrect: the lack of guarantee that every cell in a column has the same formula is worrying.
- It seems to be astonishingly hard to extract lists of candidates that satisfy certain criteria, and to format those lists in a sensible way.
- Exel spreadsheets apparently do not support dynamically updated histograms, and the making of histograms is laborious.

Fortunately, the core of the Departmental system is a database that can be manipulated with a programming language. Scaling of mark from papers was carried out on the database using this language. It was checked by a completely independent system of Perl scripts.

The transfer of data from the Exel spreadsheets to the database is a painful process because there is an absence of software that can read the Exel files and write to the database. *It must be a top priority to either fill this gap or replace the Exel spreadsheets by web interfaces ahead of next year's exercise.* Despite heroic efforts by Andy Carslaw over a sweltering weekend, corrections were still being made when the Examiners assembled for the marks meeting 5 days after the marks had all been returned by Examiners. This situation is unacceptable. Given the laboriousness of this transfer and the desire to get the earliest overview of the marks, a decision was taken to transfer unchecked marks to the database, and then after checking to make corrections directly to the database. This decision was probably a bad one.

Two BA candidates were left unclassified by the Examiners because their data were incomplete. These candidates nevertheless found their way onto the class lists initially published because Schools requires a list of candidates in alphabetical order with the letter 'I' by candidates are are not classified. This list was produced manually and the examiners who checked it must have mistaken I for 1 and had the list 'corrected' to give the candidates the classes they would have received had their data been complete. *In future these lists for Schools must be produced automatically.*

The Departmental system does not provide for scaling of Project marks. Hence once the decision had been made to scale in this way, the scaling was effected by entering by hand into the database the output from a FORTRAN program that ran on a laptop. Unfortunately, the marks were entered as the overall project mark, not just the Examiner's mark, and I did not discover this error until three weeks after the class list had been published.

When the error had been corrected, a new ranking of candidates was produced. Typically candidates had moved up or down by of order one place, and in rare cases by two places. The only border over which candidates wandered was the old 1 / 2.1 border. This was redrawn, causing two candidates to migrate upwards. This entire episode consumed a huge amount of time, both in Physics and the Proctors' office, and highlights the need for better data handling machinery.

Overall one is struck by the overwhelming superiority of the FORTRAN-based data handling system the Department had in the 1990s, to the Microsoft-based system that we now operate.

James Binney
September 12, 2006