MARKED FOR LIFE

a criticism of assessment at Universities

A. Powell and B. Butterworth

SOME QUOTES

[1] "During the course of an examination students are sometimes brought out in a state of almost total psychic collapse, shivering, unable to write, think or even to walk."—Dr. M. Conway: see section 2.

[2] "It is difficult to believe that examinations do not have some influence on the Cambridge suicides, for over half of them occurred around the examination period, and four out of five of those who were believed to be worrying over their work died in May."—Dr. A. Rook: see section 2.

[3] "There is one point in college life which is counter absolutely to the needs of adolescents—and that is examinations."—Anna Freud, quoted in Journal of the American College Health Association, 1968, p.356.

[4] "All the experimental data has shown that for a particular performance expressed in terms of an exam script, assessment by different examiners produces marks with considerable variability such that in the determination of these marks the part played by the examiner can be greater than that of the performance of the examinee."—H. Pieron, quoted in Universities Quarterly, 1967, p.300.

[5] "Pass-fail decisions at fixed proportions are . . . not the outcome but the very intention of examination processes."—Professor K. Posthumus: see section 3.

[6] "Students cannot help but see behind the friendly interest of an unassuming tutor the remorseless judgement of their Finals."—P. Marris: see section 5.

[7] "We found that it (i.e. continuous assessment) poisoned the whole teaching atmosphere."—Professor P. Edwards: see section 5.

[8] "It is sometimes claimed that students are graded by universities in the same way that eggs are graded by packing stations. This, however, is untrue. There are only two important variables determining the quality of an egg—its size and its freshness—and both of these are pretty accurately controlled by packing stations. The quality of a student's exam performance is, however, determined by a mass of variables (for example, memory, clarity and originality of thought, articulateness, luck as to which questions appear), none of which is on its own accurately expressed in the single grade awarded to each student. Thus from the point of view of accurate grading the egg gets a better deal than the student."—A. P. Ratensis.

1. INTRODUCTION

In the field of British university assessment common sense reigns; common sense, here as often, being a product of incoherent thinking and of ignorance of published empirical studies. Concerning university assessment academic staff are at least as guilty of common sense as are their students: it seems often to be assumed by staff that their own specialist learning and their past ability to succeed at certain past exams involves a knowledge of how to assess students—and thus that any effort on their part to study assessment techniques empirically is superfluous. This assumption can only give comfort so long as it is untested: few academics show signs of knowing about the great volume of evidence, compiled by academic researchers, for the conclusion that the traditional system of exams is highly unreliable. The evidence, rather more obtrusive, that exams severely affect the happiness and health of many students is usually dismissed with the claim that one's reactions to exams are useful as an index of one's reactions to crises in later life (or, often, of one's reactions to "Life"—of which the student years are not felt to be an important part).

One purpose of this booklet is to make academic staff and students aware, or more frequently aware, of certain evidence already compiled about assessment techniques of various kinds. Accordingly the studies upon which we base our conclusions are, in nearly every case, explicitly referred to, and may in most cases be easily checked. We also deal briefly with a defence of exams published recently by an eminent Cambridge don; this defence contains arguments typical of the common sense of most academic staff about exams, and is atypical only in the extent to which it concentrates such reasoning. It will be apparent that certain of our premisses in the section 'Alternatives' are anarchistic.

2. A MEDICAL ARGUMENT

That exams precipitate every year a wave of misery and psychological illness in British universities is not often disputed. And it has been shown statistically that known cases of mental unhealth among students are far more numerous around exam time than at any other stage in the year[1]. At a recent conference of the British Student Health Association it was agreed that between about 8% and 11% of all university students seek medical treatment for various kinds of exam stress[2]. There are probably many other students who suffer similar stress and do not come for medical treatment, and undoubtedly thousands of students, while suffering no important physical reaction to exams, are simply made thoroughly unhappy by them. Yet altho' these effects of exams are often commented on, those Student Health doctors who have suggested in print that university assessment may need changing in the light of its present effects are in an honourable and very small minority[3]. The unconcern of most academics about the suffering among students produced by exams is not easy to describe in restrained language[4].

Certain extreme cases of reaction to exams are summarized thus by medical writers:—"During the course of an exam students are sometimes brought out in a state of almost total psychic collapse, shivering, unable to write, think or even to walk"[5]: "Examination panic. These are the cases of students who start their papers, but get increasingly anxious or exhausted and finally leave the examination room. Sometimes they actually faint or have nosebleeds, sometimes they are overcome by headache or by migraine, but for the most it is just an increasing and overwhelming feeling of nervousness, tension, and despair, with an incapacity to remember things they previously knew. The great majority of these students have already suffered from a long period of mounting pre-exam strain"[6]: "Such (i.e. pre-exam) behaviours include all the well known symptoms, ranging from restlessness and bladder irritability to full-blown panic attacks and mania"[7]. Also, according to another medical writer, "there is reason to believe that examination stresses in some circumstances can give rise to thought disorder not immediately distinguishable from that of schizophrenia"[8]. Dr. N. Malleson has even compared students' exam-reactions to the reactions of soldiers before battle, and has suggested that the techniques for treating shell-shock can be successfully applied in cases of exam-panic[9]. In addition to these extreme

reactions, there is a mass of minor physical ailments, such as insomnia and anorexia, precipitated by exams which comes annually to the notice of Student Health centres. Considered together with this range of reactions, the various kinds of fear which afflict most students before exams constitute a strong argument for exams being immoral—unless it can be shown that exams in other ways contribute greatly to the well-being of society and that they cannot be replaced in this function by anything more humane.

It should not need saying that those students who suffer most from these reactions to exams under-represent their abilities in the exam-room. (It seems that the effect of anxiety on exam performance, plotted graphically, forms a curve; up to a point, the more anxious the candidate the better, on average, he or she performs: after that point, the more anxiety there is the worse, on average, the performance[10].) There is an idea, often expressed or implied in academic circles, that exams, while unfair to the over-anxious, are fair to everyone else. This is wrong for at least two reasons. First, and obviously, because in the various competitions for jobs and grants in which exam-grades are used, to give an unfair disadvantage to some is necessarily to give an unfair advantage to others. Second, for a reason more complex and more rarely understood: research has shown recently that many examiners operate, consciously or not, a quota system when awarding grades, independent of the actual standard of the group of scripts marked[11]. Different examiners have different quotas — e.g. their quota of Firsts may be anything between 0% and 14%—but whatever their personal quota, it is adhered to with a regularity which transcends variation from year to year in group performances. Now any fixed quota system means that the more people in your group who do worse than you, the better grade you will tend to get. Thus through the under-performance of very anxious exam-candidates, many others every year will find themselves in grades higher than they would otherwise have achieved. All serious competitions are unsavoury, but this competition, which enables some students to profit automatically from the impairment of a minority within their own group, is especially unjust and unpleasant.

It is not true that those students who get most upset by exams are usually those who have "done no work". Altho' students known to be suffering from exam stress do on average rather worse than the generality (a fact explicable by their over-anxiety itself), they are by no means

concentrated around or below the pass/fail line[12]. And even if it were true that most sufferers from stress at exam-time were people who had done less work than average, that would not make inaccuracy in their exam results any more tolerable—unless it were desired to punish such people by giving them an exam-grade which under-represented their true achievement on a course. There is some evidence that two special groups suffer particularly from — and so under-represent themselves because of — anxiety about exams: women and overseas students[13]. In a study of Manchester students, almost twice as many women as men proportionally were found to present symptoms of exam stress. The figures for overseas students who suffer in a similar way are not so extreme but are still high; part of the explanation of this is probably that in many cases the social pressure on overseas students to succeed in exams is unusually high — much money and hope having often been invested in sending them to a British university[14].

Certainly attempts are made to get examiners to take into account evidence of psychological impairment when they grade candidates; of the 138 medical reports sent to examiners in University College, London in 1970/71, most were concerned with impairment through exam-stress. Yet even assuming that examiners are often influenced by such reports (a generous assumption, since many academics think that over-anxiety *should* be reflected in exam marks[15]), it is unlikely that such reports are made except in the more extreme cases of impairment, and in any case there is no way for the examiner to accurately measure how an individual would have performed if he or she had not been psychologically impaired.

One of the skills crucial for succeeding at exams, that of solving complex problems, is very badly affected by nervousness in the exam room (v. supra, note 10). It is widely agreed that an over-nervous candidate typically reacts to an exam question by reproducing a great deal of—and often too much—information, with far too little logical connection. Often an examiner is faced with a script which does little more than assemble information from which the particular question could have been answered, had the candidate been less impaired by anxiety. And yet problem-solving—involving the grasp of logic and scientific method—is probably the most important and lasting academic quality which a university can develop in people; ten years after graduating, when you probably remember so little about your degree subject

that you could not pass 'A' Level in it without revision, you probably still have most of the logical and methodical ability which you possessed at university. Yet this ability, highly important because of its permanence and wide applicability, can never be fairly tested by exams because of the high extent to which its expression is affected in many people by exam stress.

"But examinations reflect the sort of situation you're going to meet in Life. Employers are entitled to know if a man is going to crack up under pressure." This type of answer is the only one commonly made to the criticism that nervous impairment on its own destroys the accuracy of exams[16], and its implausibility ranks it with those propositions which, in Aristotle's phrase, "no one would assert unless he were defending a thesis". For one thing, as is well known, it very rarely occurs in life outside the exam room that one has to take very elaborate and important decisions while isolated from all books, papers and consultation with other people. Also, most people who have taken their Finals will never again face a period as short as the exam period knowing that by their performance in it, tho' that performance may be no worse than mediocre and may be well below their own usual standard, they may be for ever disqualified from the career they want to follow, and may even be branded as a general failure. The process of selection by which most of present society works consists largely of continuous assessment[17]; in Stengel's words, "bishops, admirals, judges, professors, not to speak of Cabinet Ministers, are invariably elected without examinations"[18].

However, granting for a moment that situations like those of exams did occur in the lives of most graduates, it would still have to be admitted that they were very unusual situations. Why should the sole important paper qualification that most universities give be a measure of performance only in such very unusual situations? Even assuming that exam conditions corresponded with other situations, it would still be irrational for universities not to attempt to give a measure of performance under normal and far less anxious conditions[19].

In any case it is now accepted within the medical profession that people who are over-anxious in one crisis are not necessarily over-anxious in all crises[20]. How anxious one is in different situations varies very much with the nature of one's circumstances and psyche at the

time. Thus the student whose parents expect a good result, or who has been financed at great sacrifice by others, or who fears that he/she is inferior and that an exam result will confirm the fear, is under pressures which may well not be duplicated in later life. It cannot, therefore, be argued that over-anxiety in the face of exams means over-anxiety at other crises in one's career. Conversely, success at exams is compatible with being badly impaired by over-anxiety at other crises. It seems that the tendency of exams to inspire panic, over-anxiety and misery may require a more ingenious justification than has yet been evolved.

NOTE: I wish to record my gratitude to Dr. C. J. Lucas, of the U.C.L. Student Health Service, who contributed invaluable information and bibliography for the above section.

It has long been widely agreed that male[21] students in England and Wales are, as a whole, far more suicide-prone than their non-student contemporaries[22]. Suggested explanations of this phenomenon have been few and very cautiously expressed[23]. Twelve years ago an eminent medical writer, A. Rook, K.B.E., F.R.C.P., published an analysis of suicides at Cambridge university occurring between 1948 and 1958, and concluded that "it is difficult to believe that exams do not have some influence on the Cambridge suicides, for over half of them occurred around the exam period, and four out of five of those who were believed to be worrying over their work died in May"[24]. It seems that this suggestion of a causal link between exams and suicide provoked a swift reaction at Cambridge; within a year the Statistician to the Medical School of that university, R. G. Carpenter, published a similar study which, while it did not dispute Rook's figures for the period 1948-58, claimed that over the longer period 1923-58 the proportion of suicides at Cambridge which occurred in the third term (the term when exams happen) — 43.3% — was not significantly higher than the proportion of suicides among a control group of non-student males at corresponding times over the same period — 34.5%[25]. It should be noticed, however, that a considerably higher — if not a "significantly" higher — proportion of the Cambridge suicides did happen during the summer term than in the non-student control group (43.3% : 34.5%).

In spite of the revelation by these two studies that there was over a long period a high concentration of suicides at Cambridge during the summer term, there has never been published a similar study of the timing of suicides at any other British university. Whether suicides at university have in general been to any extent abnormally concentrated in this term is simply not open to the layman to discover (at least, without writing to every Student Health Department in the country)[26]. The failure of the medical profession to provide such a general study is all the more remarkable because Carpenter himself pointed out that the proportion of summer term suicides at Cambridge between 1948 and I958 was much higher than in previous years, and wrote that this "might suggest that there may have been a change in the time of year at which cases tend to occur"[27]. Whether this apparent change has been sustained

in the period 1958 to date is very important to know for anyone concerned to test whether exams have an effect on the number of suicides.

It is not my contention that exams generally precipitate suicides in the universities of England and Wales. However, the only proper studies published on the subject suggest that exams at Cambridge between 1948 and 1958 have had this effect, and because it remains a possibility that this effect obtains more generally the failure of the medical profession to publish a thorough study of the question is deplorable.

3. THE FAIRNESS OF EXAM GRADING

"In spite of all its faults, the examination system remains a more rigorous test of quality than any possible substitute." — an unnamed don, quoted by D. A. Allen[28].

Students often complain about particular injustices of the final examination system, especially when, apparently by chance, they do worse than expected. Academics occasionally do, notably if a prize student finally fails to fulfil his mentor's hopes. But almost no academic, and precious few students, query the fundamental justice of this assessment system. They believe that if a student gets a First, he is clever and knowledgeable; and a student, if he gets a First, will consider himself clever and knowledgeable. If he gets a Lower Second or a Third, the general view, including his own, is that that's all he's worth. It's very difficult for the student, in any case, to tell whether justice has been done, since no examiner's account of his work is made available to him: and if he does worse than he had hoped, he will most likely put it down to what he imagines were inadequate answers on papers one to five, or whatever. This is perhaps not surprising, given that the student spends three to four years in pursuit of a good degree, and would therefore be loth to admit that he has wasted some five per cent of his expected life-span in doing so. And academics are permitted to do the job they do because they have done well in their final exams. They too would not care, very often, to entertain the idea that they have achieved this relatively coveted job largely by chance, and that they may have thereby unfairly denied some more worthy candidate the position. So we find them making unsupported claims about the exam system, like the one quoted at the beginning of this section.

Nevertheless, there is a body of experimental evidence stretching back to at least 1930 which demonstrates quite clearly that chance, and a variety of irrelevant factors, are ingredients, probably the main ingredients, of finals grades. And it is a measure of the complacency and self esteem of the academic establishment, and to some extent of students also, that this evidence has been consistently ignored.

To assess the importance of irrelevant ingredients in finals grades—that is, to assess the validity of exams—is to assess whether exams test

what they set out to test and nothing else. One trouble is that there is very little agreement among academics as to what exams are meant to test. Jahoda and her colleagues at the university of Sussex have examined some of the assumptions underlying the use of exams, and list as the first of these the belief that exams test "knowledge and quality of mind"[29]. Daniels and Schouten, writing a report for the Council of Europe on university screening methods, see exams as primarily asking the question "Can a student at a stipulated time give proof of a certain well-defined knowledge?"[30]. Others trying to define the central intellectual purpose of exams come up with similar, and more or less vague, formulations. The unnamed don already quoted sees further purpose:

> "There is strain and tension, there is working against time, and the student has to use his intelligence and his accumulated knowledge in the way that he will be called upon to use it for the rest of his life. Examinations are thus a psychological and moral test as well as an intellectual one."[31] We may for the time being regard this as a minority view.

In one sense, the main purpose of exams is social and not academic, and it is so obvious as to be often overlooked. Exams are meant to provide a standard of comparability from one subject to another, from one year to another, and also within a year, subject and university, between one candidate and another. These candidates will, presumably, be answering different questions in different ways. But one must be able to say that a student of history who obtained a First at Sheffield in 1958 is of the same standard as a student of chemistry at U.C.L. who got a First in 1971. And of course, one must be able to say that this year's First in philosophy at U.C.L. is of a better standard than this year's Second in philosophy. On these grades depend salaries, job opportunities and, to some extent, how one regards oneself and is regarded by one's peers. Yet it is not clear what these grades are meant to measure. Among the things exams are clearly not meant to measure are: what year it is, what subject is being examined, and which university the candidate is at. If they measure any of these things, they are failing in their purpose. However, it is equally clear that they do measure these things.

The Robbins report of 1963 showed that exam candidates in British universities are classed in proportions that tend to remain the same within particular universities and faculties[32]. This observed tendency is

consistent with the findings of another research project into exam grading—the conclusion of which is profoundly damaging to the claim that exam grades are valid indicators of academic quality. A Dutch educationalist, Posthumus, recently found, after studying exam marking, that the drawing of pass-fail lines was often being done according to a fixed proportion, and independently of variation from year to year in the quality of the groups of scripts marked[33]. This fixed proportion varied with different faculties and with different examiners. In Posthumus' words, "Pass-fail decisions at fixed percentages are, in fact, not the outcome but the very intention of examination processes."[34]. An effect of such a fixed percentage system is to make your chance of failing an exam far higher if you happen to be in a better-than-average group than if you were in a worse-than-average group—even tho' your own exam scripts did not vary at all. Also, your chances of failing are obviously higher if you are examined by a faculty which operates a high fixed proportion of failures than if your faculty has a low fixed proportion of failures. Almost certainly this pre-setting of the pass-fail proportions applies equally to the other grades within exams; this would best explain the stability within grade structures clearly shown by the Robbins report.

If all grade proportions are thus determined in advance of marking, it follows that your chance of getting a First or an Upper Second is also less, the better the other students in your group are. Whatever your grade is, it will at best only reflect your rank order within your group. Even if every faculty had the same fixed proportion of Firsts and Upper Seconds (which they do not), it would be absurd to claim that a First or Upper Second in one faculty was equivalent to a First or Upper Second in another, unless there were evidence that the standard of candidates within the faculties compared was the same. Such evidence has never been produced by different universities even within the same subject: and within different subjects it is hard to see how achievement could ever be made commensurable. An exam grade of the traditional kind does not, in all probability, measure you against the generality of other students, but merely measures you against the people in your group. Since the quality of your group is an unmeasured variable, the value of any such exam grade is very small.

There are also considerable, and consistent, differences in the proportion of Firsts and other grades awarded in different subjects:

TABLE 1

(showing proportions of Firsts among students graduating in various subjects in 1962)

Social studies	3%
History	4%
Theology	5%
Philosophy	5%
English	6%
Modern languages	7%
Geology	9%
Classics	12%
Chemistry	12%
Physics	12%
Maths	14%

(Robbins Report, 1963, Appendix 2A, Part 4)

— These figures are consistent with the findings of Dale in 1959 that in British universities that year 10% of students graduating in science got Firsts, while only 5% of those graduating in arts subjects did so; Dale, Universities Quarterly, 1959.

Science and technology students get between two and three times as many Firsts in proportion as do arts and social science students. There is no support for the hypothesis that this is because the former students are better. Indeed, the qualifications of science students, in terms of 'A' Levels, are markedly worse than those of arts and social science students. This might simply point to a discrepancy between the predictive value of arts and science 'A' Levels in terms of university achievement. However, if one looks at the proportion of the other grades, a different explanation presents itself. Science students also get far more Thirds and Fails in proportion than do arts students. In statistical terms, the spread of marks is much greater in science. This suggests that science exams are marked in a different way. Entwhistle has pointed to this difference, and remarked that whereas in science exams it is possible to get close to 100% for a question, and also close to 0%, in arts exams the range of marks tends to be restricted between 25% and 75%.[35] Science grades and arts grades are thus not strictly comparable. University exams do in fact test what subject one is taking.

Other factors can be shown to influence grading. Cox reports an experiment by Kandel into the effects of handwriting on exam scores: "Markers were instructed to give so many per cent for handwriting, and no more. When the essays were typed out, however, and re-marked, it was found that in fact handwriting had affected the marking of the other aspects of their answers"[36]. The preceding section, 'A Medical Argument', shows that the way a student is affected by anxiety will in turn affect his grades, and may indeed incapacitate him so seriously that he will be unable to take the exam. These are just some of the factors which university exams test: this disproves (and may eventually even eradicate) the traditional, complacent view that these exams provide a strict standard of comparison of the academic merit of candidates. The factor which is most frequently held to lessen the value of grading has not yet even been mentioned: luck as to which questions the candidate finds himself faced with on that fateful day in June. But see below.

More shameful and surprising even than the clear invalidity of the exam system is its unreliability. There has been a number of convincing experimental studies that exams are highly unreliable, dating back as far as 1888[37]. Cox, in an excellent article already cited, reviews much of the work in this area. He describes the formation of the classic study in the unreliability of exams[38]:

> "There was, then, quite a lot of rather tentative research available when the Carnegie Corporation of New York gave funds to support an international conference on examinations. At this conference, national commissions were set up to produce evidence for discussion at later conferences. The English commission of eight members contained three very eminent psychologists, Cyril Burt, C. E. Spearman, and Godfrey Thomson. The experiments carried out by this commission were extremely rigorous, and every attempt was made to select only very experienced examiners and avoid any artificiality. The scripts used for marking were not specially written ones but were taken from actual examinations. The results are most fully set out in 'The Marks of Examiners' by Hartog and Rhodes with Cyril Burt (1936)."

Hartog and Rhodes report on the reliability of marks for three subjects: English, history and maths. Fifty English papers from college entrance exams, in which the candidate had to select one from four questions,

were used as experimental material. Five examiners marked these papers out of 100 marks. The mean range for the five examiners on each paper was 19.6 marks; the greatest range was 36 and the smallest was 7. In history, four university honours papers were selected, taken by eighteen candidates. Three of the papers were marked by ten examiners and the other by five, using 24 grades in all. The mean range per candidate was 7 grades on the first paper, 11 on the second, 10 on the third and 9 on the fourth. Ranges as large as 18 were found. The correlation between examiners varied from −0.41 to +0.85, with an average of only +0.44. For maths, a university honours paper was used; there were twenty-three candidates and twelve questions. Full marks could be obtained by doing six questions. Six examiners marked each candidate out of 300. The mean range per candidate was 34.7, with ranges marked as high as 64 and as low as 17. In all subjects the average marks given by any one examiner did not differ very much from that of any other examiner: thus their high ranges of marks per candidate did not depend on systematic differences between particular examiners—i.e. it was not that one examiner marked consistently low and another consistently high.

In other studies, examiners were asked to mark a paper and then to re-mark it at some later date. In all these studies, the difference between what an examiner gave to a candidate on one occasion, and what he gave the same candidate on another, is astounding. In 1930 Eells had 61 teachers mark and remark two history and two geography essays at an interval of eleven weeks[39]. The average correlation between each teacher's marks on the two occasions varied between 0.25 for one essay and 0.39 for another, with an average of 0.365. For an apparently more factual subject, medicine, Bull had an examiner re-mark thirty scripts after an interval of several weeks. The average mark on each occasion was very similar, but the correlation between the marks he gave a candidate on one occasion as compared with the mark on the second marking was only 0.28, which fails to differ significantly from chance[40]. In other words, if the marks had been allotted randomly there would have been a better than one in ten chance that this was the correlation achieved! An analysis of the actual marks reveals that with a pass mark set at 50%, eight out of the thirty candidates marked by this examiner would have changed from pass to fail or vice versa on being remarked by him. Bull cites many other examiners who failed to better a mark-remark correlation of 0.5.

The highest correlation between two examiners—including between one examiner on two occasions—that can reasonably be expected is about 0.85. Daniels and Schouten point out that even with a correlation as high as this, if one takes a pass mark of about 50% and, say, a hundred candidates, of some sixty four failed by one examiner about ten would still get a pass from the other: and of the thirty-six passed by the first examiner about six would be failed by the other[41].

Thus even with this maximum correlation about sixteen per cent of candidates are still subject to a pass-fail difference between examiners. And of course with lower correlations, which is what we usually get in practice, this proportion might be as high as 70% of all candidates. And it does not help much to average the marks given by the different examiners. Suppose there is a "true" mark for a paper: if one examiner wildly exceeds that mark, in order for that mark to be attained the second examiner must equally wildly undermark. If the latter does not, the average mark is still higher than it should be. And what tends to happen as one averages more examiners per candidate is that the mean scores for each candidate get closer and closer together, making the separation into grades even more arbitrary than it is at the moment. In any case, increasing the number of examiners to improve reliability is hardly practicable, since, as Bull pointed out, reliability increases not directly as the number of examiners but approximately as the square-root of their number[42].

The evidence, therefore, can only be accounted for by assuming that exam grading is not reliable, and that its unreliability is such that your final exam grade will depend to a very large extent on luck. This is not an occasional phenomenon, but one which seems to permeate the whole exam system.

The grading is not only unreliable in the sense that examiners do not agree among themselves, or agree with their own previous markings. It has not been established that a given candidate will do equally well on the same test on different occasions. For exams to be reliable, and indeed for them to be valid, a candidate's performance should vary little between occasions, otherwise what is tested is the occasion itself. Clearly, one cannot give exactly the same test to a candidate on two occasions,

since he will know the second time what the questions will be. It is therefore difficult to test this sort of reliability. If one sets different questions, these particular questions might be better or worse prepared than the previous questions, and so the tests will not be strictly comparable. But this is precisely analogous to what happens in the real exam situation: it is more or less a question of luck whether the candidate has prepared the particular questions he finds on the exam paper. He cannot prepare the whole field equally, and therefore he cannot be fairly compared with other candidates, since it is a matter of mere chance whether they are equally prepared for the questions they find, even supposing they have all worked equally hard for three years.

Attempts to make exams more reliable have tried systems where the candidate must answer more questions per paper, so that a greater proportion of the field must be covered by each candidate. I do not know of any research where this technique has been used to test improvements in intracandidate reliability, but it has been used to test interexaminer reliability. Bull, investigating exams in medicine, tested the changes in marker reliability, both between-marker reliability and mark-remark reliability for each examiner, with changes in the number of questions the candidates had to answer in a three hour finals paper. There were four experimental conditions: 4 answers in 3 hours, 8 answers in 3 hours, 16 and 32 answers in three hours[43]. Mark-remark reliability improved from about 40% for 4 to 67% for 32 answers for one of three markers, and by similar proportions for the other two. Between-marker correlations also improved somewhat. But the most staggering improvement appeared when the analysis of variance was calculated. The percentage variance due to students was 71% on the 16-answer paper, but a tiny 19% on the 4-answer paper! This means, in simple terms, that 81% of the variance in marks is due to the examiners—that is, it is due to factors unconnected with the student.

However, trying to improve either reliability between markers or intracandidate reliability by increasing the number of questions changes the nature of the exam. The more questions the candidate has to answer, the shorter these answers must be. They will therefore not test the ability to elaborate supported argument, which traditional exams purport to test.

The extreme case of increasing the number of answers required is the multiple choice test. Here one can sample a very large proportion of the syllabus, and get 100% reliability, since machines can mark the answer sheets. However, for the same reasons that reliability increases, validity seems to decrease proportionally. It is usually claimed by academics that altho' mere memory of facts is necessary for finals, it is how they are put together in answer to particular questions that is really being tested. Now as you increase the number of questions you decrease the opportunity to put facts together in this way, and multiple choice papers become little more than a test of memory. Moreover, a multiple choice paper assumes that the examiners, before they mark, know the right answer, or all the right answers, to every question, since it is they who pose all the alternative answers from which candidates have to choose. There is no opportunity for a candidate to gain credit by giving a novel answer—something which is at least possible under the present system.

It is not clear why examiners should differ in the way they do, but it is clear that differences in general standards, or in understanding what is required of them, are not the reason. In the Hartog and Rhodes study the mean mark given by different examiners varied very little—it was not that examiner A always marked low and examiner B always marked high. Stalnaker had English teachers mark essays by school seniors, having first extensively discussed with them the standard at which to mark, and indeed having had practice sessions in marking where the results were carefully analysed. Nonetheless, the overall average correlation between marking and re-marking was still as low as 0.55[44]. Cast had markers mark and re-mark essays in four different ways. The overall correlation was about 0.492, but the correlation for the condition where analytic marking techniques were used to establish standards dropped to 0.485[45]. Both of these studies used school essays. One would imagine even greater discrepancies for university essays.

The real problem seems to be that all the kinds of assessment techniques that have been discussed above try to put all the different abilities which students bring to examinations into one basket—the grade. It may be the case that a hardworking student who has a good grasp of his subject will get a particular grade, say Upper Second, and that a more imaginative but less thorough student will get the same grade. Decisions affecting the careers of both students will be made on the basis of this

grade, yet the grade serves not as genuine characterisation of the abilities of these people, but rather to disguise vital differences. It is therefore not surprising that the evidence presented in this paper shows that exam grades are unreliable, since one examiner may be looking for one quality and a second examiner, or the same examiner on a different occasion, may be looking for another quality. The fact that grade averages between examiners can be similar does not mean that they are both working on the same assumptions, but rather that the qualities they are looking for may be distributed in similar ways over the whole population of candidates that they are examining. But this of course needs further research, before being assertable with any confidence. And it goes without saying that exam grades are invalid. The grading system assumes that there is one ability which has a continuous distribution, and that this continuum can be neatly and nonarbitrarily divided into four or five meaningful categories. This point only has to be stated for its absurdity to be evident.

The N.U.S. in an executive report, after a cursory examination of the exam system, state "The NUS believes that only by a combination of examining methods can the majority of students *feel* that they are assured a fair deal. We do not assert that all students will do better when a mixed system of assessment (i.e. objective tests, traditional exams, open-book exams, continuous assessment etc.—Ed.) is introduced, but at least each student will *feel* that he has a chance to do justice to himself."[46] Now it is obvious from this quotation (and from the rest of the pamphlet) that the NUS executive do not seriously question the assessment system as such. They appear interested only in how the student "feels", and feeling one has a fair deal is not the same as having it. They even entertain the possibility that every student could "do better", which merely shows that they have totally failed to understand how the system works, since as has been pointed out above "decisions at fixed percentages are, in fact, not the outcome but the very intention of the examination process." For each student to "do better", it is only necessary to change these percentages. The real criticism of the NUS solution is that to put the results of "mixed assessment" into the same four or five baskets as the results of traditional assessment, is as absurd, if not more so.

It is clear that the unitary grading system is a waste of time and money which could be better spent educating people. No one has

calculated the cost of university exams; the building or use of special halls, the printing of papers, the payment of invigilators, the cost of academic man-hours in setting and marking the papers etc. etc. must use up a sizable proportion of any college's budget. It would be worth knowing exactly what proportion, and considering how these resources might otherwise be employed. But these exams are worse than just a waste. They purport to give candidates, and the rest of the world, a true and objective account of students' abilities. It is clear that they do not, and there is no evidence that they ever could. If members of the academic establishment are made aware of the evidence for the invalidity and unreliability of university exams, and still persist in claiming that exam grades are an accurate account of academic abilities, they will probably be guilty either of extreme folly or of a confidence trick.

4. A CONSERVATIVE VOICE (radically answered)

To be successful, a small radical group needs to justify itself constantly with argument. Those in power, however, usually do not. For them pure force can achieve what others need a mass of reasoning to bring about—a fact which Louis XIV used to acknowledge by stamping on his cannons the motto that they were "A king's last argument" ("Ultima ratio regum"). This characteristic of power is doubtless one reason why so few reasoned defences of the present exam system exist in print: another reason is obviously the inherent difficulty of defending exams in the face of empirical studies such as those we have cited. However, college authorities will sometimes give short, oral "common sense" defences of the exam system, and recently there was published a justification of exams which has very much in common with these oral defences[47]. It was made by J. Chadwick, Fellow of Downing College Cambridge, an examiner and a scholar renowned in his own field for his role in the decipherment of the ancient Greek Linear B script. Chadwick's defence of exams is complacent, self-contradictory, superficial and seemingly uninformed by any knowledge of published scientific studies of the exam problem. Being in these ways typical of academics' arguments for exams, it has been selected for special treatment in this section.

Chadwick begins, "I think the great merit of the Cambridge (in almost every respect one might say the British) examination system is that it is not only fair, but can be seen to be fair."[48]. It is not clear how this statement is compatible with the mass of published evidence that not only do different examiners give widely differing marks to the same scripts, but that the same examiner on re-marking scripts will usually give very diverse marks. Indeed, it is not clear that Chadwick is even aware of this evidence. He goes on, "The problem of fairness . . . rules out the proposal for 'continuous assessment'. Who is to do the assessing? Directors of Studies? Supervisors? I must admit that *I* would not have such unbounded confidence in my colleagues or myself." This time it happens that there is some evidence to support Chadwick's claim; oral exams have been shown to be even less reliable than written ones[49], and continuous assessment by staff in the student's own department would inevitably employ a judgment on the student's oral performances. This admission by Chadwick that academic staff are highly subjective and

unreliable in assessing their own pupils seems, therefore, quite correct and laudable. Its effect, however, is rather spoilt by Chadwick's claiming, on the following page[50], that "Fairness is ensured for the Ph.D. candidate by his supervisor and by the oral examination." So it appears, after all, that Chadwick is unaware of the evidence for oral exams being very unreliable. Nor is it made clear how a supervisor, whose subjectivity is admitted to be so great as to disqualify him from assessing his First Degree students, is able to "ensure fairness" for his own Ph.D. students. (A Ph.D. student is allowed only one supervisor at a time, and at London university—tho' not at Cambridge—that supervisor is also one of the student's two examiners. Since for all Ph.D. students the reference provided by this single supervisor may be crucial for securing a first academic teaching job, the practice of Ph.D. students having only a single supervisor does not avoid subjectivity—it enthrones it.)

Chadwick also commits one of the most familiar fallacies among arguments for exams. This is the claim that while one examiner may mark unfairly (itself a telling admission), the presence of a second examiner will rectify the injustice: "Under the Tripos system of double marking and examiners' meetings the idiosyncrasies of dons tend to cancel out, and the result is fair (at times over-generous) to all."[51]. Now assuming, to be generous to the system, that only on one script out of ten does an examiner deviate far from a "true" assessment. Then, for that script to be marked fairly, since examiners compromise, it will be necessary for a second examiner to deviate to an equal extent in the opposite direction. The chance of this happening will be at best one in 200, and there will, of course, be an equal chance that both examiners will deviate in the same direction. Certainly a "true" mark may be produced accidentally through symmetrical error, but what is far likelier to happen is that one examiner will produce something close to a "true" mark and the other a grossly deviant one. The resulting compromise between the examiners will not remove injustice, but will merely halve it.

Consider, too, the latter part of the sentence quoted: ". . . the result [of exam markings] is fair (at times over-generous) to all." Now within the same group of candidates it is impossible for some to be treated fairly and others over-generously. For the over-generous treatment of some will reflect unfairly on the relative merits of the others. But perhaps Chadwick means that when over-generosity occurs, it occurs for all those marked under the one exam? But if this happens, how will the

result be fair to those who are not marked over-generously, in different universities, faculties and (at Cambridge) in different groups within the same Tripos? One of the prime uses of exam grades is to enable people in different universities, faculties and groups to compete in the same competition for various jobs and grants. Thus to be over-generous to one group is to be unfair to all those who are not over-generously treated. Chadwick, it seems, must choose either to claim that exam results are fair to all or that they are over-generous to some and thus unfair to others. To claim both is self-contradictory.

We are also treated to a variety of the "Life's like that" argument about exams. Chadwick says about the stress of Finals, "There are of course people who go to pieces under that sort of pressure; which is another way of saying that a First is not simply a certificate of academic brilliance, and most employers would like to be warned if a prospective employee can only be trusted provided he has plenty of time and no pressures on him. In most professions, life is not like that."[52] Nor, however, is life in most professions ever like a Final exam. But see above, section 2.

Finally there is an interesting revelation of the conservatism of our author; he objects to giving students more scope in choosing their own subjects for examination on the grounds that this "runs into administrative difficulties"[53]. The only justification he gives for this claim is that a wider variety of subjects would lead to clashes in the lecture timetable. This itself might seem a small thing to be allowed to restrict freedom of study. But when it is realised that nearly all lectures are anyway better replaced by printed handouts, Chadwick's position is exposed as no better than arbitrary. In short, Dr. Chadwick's opposition to the reform of the assessment system seems to owe its force less to reasoned argument than to his institutional power—ultima ratio regum.

5. SOCIAL EFFECTS OF GRADING

The jealousy which many students show near exam time about their own knowledge, understanding and theories is well known[54]. Very probably this jealousy results from a feeling that if one's script "stands out from" others it is likelier than otherwise to get a high mark: numerous students almost certainly feel that the worse their class mates do at exams, the better is their own chance of getting a high grade. And recent research has shown that such a feeling is in many cases justified[55]. But even if the feeling were not well-founded, its very existence would still have important social consequences. The belief that one's success may depend on the failure or mediocrity of other people in one's group must vitiate a university as a place of learning. Students in a group are often facing the same novel problems of learning as each other, and with very similar academic histories: for this reason they will often have a better insight into each other's difficulties and interests than academic staff can have, since it may be anything up to 45 years since the staff went through a similar intellectual stage. Students thus have a great deal which only they can teach each other, and any sense of direct competition which inhibits this mutual teaching severely damages the process of education. Continuous assessment might create such a sense to an even greater extent than exams do at present, because under continuous assessment far more of one's work at university would be used as a basis of personal evaluation; even the weekly essay might come to be thought of as part of a competition against one's fellow-students.

Several academics have written of the tension between staff and students which results from the exam system[56]. As Marris puts it, "Students cannot help but see behind the friendly interest of an unassuming tutor the remorseless judgement of their Finals."[57] And many medical teachers are reported to welcome the fact that their students are examined by an outside body, because they feel that their relations with their students are thus not spoilt by their having an examining role[58]. Continuous assessment seems to have an even worse affect in this area. At the universities of Sussex and Essex, where forms of continuous assessment have been tried, staff-student relations have been felt to worsen as a result, and a professor at Essex is reported as having said of continuous assessment that ". . . we found it poisoned the whole teaching atmosphere"[59].

"But you need grades to make people work." There is very little evidence for the general truth or untruth of this claim. A questionnaire administered to American students drew the response that under their grading system (which assigns grades far more frequently than in Britain) a high grade in an interesting subject encouraged work, and in a boring subject discouraged it. Only low grades, it was felt, had the general effect of encouraging work[60]. Most claims by students as to how much work they would do if grading were abolished are of little value; very few students, if any, have real (rather than imaginative) experience of their own reactions to an education system which avoided grading altogether. And even if it were true that the removal of grading now in British universities would at first reduce standards of work[61], this would not mean that the absence of grading need always have this effect. It might be that the use of grading from secondary school up had caused scholastic work to be looked on above all as the medium of an unusually stressful and public competition—and thus as something inherently unpleasant. How work at schools and universities would be regarded by students if it were not used constantly as a basis for assigning opportunities and disqualifications, praise and humiliation, remains for experiment to show.

By discouraging students from co-operating with each other the assessment system inhibits a prime virtue of civilised society—that of mutual aid. By isolating people from each other in a highly formative stage in their lives, and encouraging them to regard their work as a private and measurable achievement, it enforces or reinforces the view that different people deserve different rewards in life. If it were made clear that we owe a large (tho' not a precisely measurable) proportion of our knowledge and ideas to the people around us in society, and that our own contribution to society similarly defies measurement, many more people than now might wonder why our wages and job opportunities should be precisely differentiated from those of other people. The process of grading at universities seems, therefore, to be not only an attempt to select people for different strata in society, but also, in its effect, to be a psychological preparation for accepting a stratified society.

6. ALTERNATIVES

People often assume that to desire the abolition of exams is to desire the introduction of continuous assessment into grades. And certainly some forms of continuous assessment seem to be much less unpleasant for the student than the traditional exam; in at least one British university where continuous assessment has been used to determine final grades the number of students seeking medical treatment for stress has been "dramatically reduced"[62]. But several arguments we have already adduced against grading by traditional exam apply also to grading by continuous assessment. The allotting of a single grade to each student can never describe that student's various intellectual qualities and performances in different areas of a course. The setting of grade lines must always be arbitrary, even if it could be standardised, and there is no reason to suppose that assessors will be any more consistent in the marks they give to work over a long period than they now are in marking traditional exam scripts. As has already been pointed out (in section 4), since continuous assessment would probably embody some evaluation (conscious or not) of students' oral performances, it might be even more unreliable than assessment by traditional exam—since research has so far indicated that examiners are even more subjective in their marking of oral than of written work[19]. The seemingly bad effect of continuous assessment on staff-student relations was illustrated in the previous section of this booklet. And that form of continuous assessment which would impose a series of petty tests on students should be anathema to any educationalist, because it would heighten the sense of competition within student groups, and even more than at present would subordinate study to the purpose of passing tests.

The abolition of all grading apart from the pass/fail division seems quite possible within our present structure of society. Such a system already operates in the awarding of Ph.D.'s, and also in medicine at the level of the final M.B. Assuming that the pass/fail line for first degrees stayed in or near its present area, this two-grade system would have the advantage of cutting down isolation and competition among students, and of encouraging co-operative work; most students would expect to get the same grade, namely 'pass', and far fewer people than now would feel they had a direct interest in their fellow students getting low marks. However, this system would still have important faults. The fixing of

the pass/fail line would be as arbitrary and unreliable a process as the fixing of any other grade. And the effect on the feelings, the confidence, and so on the achievements of anyone marked 'failure' is not pleasant to contemplate.

The university system most worthy of experimental adoption, as we think, is that involving the abolition of all grades, and even of the pass/fail line. Syllabuses thus could atrophy, since their main purpose is to standardise learning in order to make assessment more rational. Lecturers would then be free to teach anything that students wanted to hear and talk about, and students would be able to transcend the traditional categories of learning. Certificates could inform anybody interested that so-and-so had attended whatever courses and had produced a certain volume of work, and beyond that each person would have to be judged purely on their merits for whatever job they applied for. It would of course be open to students, when applying for any job, to present samples of any relevant work they had done at university, just as, at present, artists use portfolios of their work.

Under present society, such a system might have the effect of creating a number of special job-entrance exams, like those at present existing for entry into the Civil Service. But such exams would still leave universities largely free of the sense of competition which currently permeates them, and failure at such exams would not seem anything like so general and humiliating a mark of incompetence as does failure in traditional exams. Consistently applied, our principles would involve the removal of university entrance tests, G.C.E.'s, C.S.E.'s, 11+'s, and all other devices of formal education for the general disqualification of people. To remove thus the differential opportunity system from formal education would not only improve relations between students, but also between students and staff; the latter would cease to be seen and resented as having the awesome role of examiners and of ministers to exams. And we could all get on with something more useful in the summer term. Far more importantly, free from formal competition and differential opportunities, the education process would help spread among young people a rejection of the system of differential rewards which awaits them in later life. After having co-operated with each other for years as, in many senses, equals, without formal rank and without special rewards, many more young people might come to see the unequal division of wealth and of unpleasant work in society as the divisive and unnecessary evil that it is.

"But surely, while society has pleasant and unpleasant, well and badly paid jobs, there is always going to be some competition among students to get them, even if you do remove all grading from formal education?" Quite right, and you may think that the abolition of *that* social fact should be the next job.

7. ON ACTION

Grading will not wither away—it needs to be abolished, and most of the pressure to abolish it will come from students. But it would be badly wrong to run a campaign against examination as if it were a campaign against all staff; there is a small minority of staff already in favour of abolishing grading at universities (see, for example, Universities Quarterly, 1967, p.351). This minority is expandable, and will, for several and mostly unrespectable reasons, have a force per caput that students cannot hope to have. Also, as we have found, sympathetic staff can supply valuable insights into how the opposition of their less enlightened colleagues may be met. Students should carefully resist the encouragement of university administrators to "discuss exams primarily at departmental level"; as administrators well know, at departmental meetings academic backwoodsmen are most highly represented and effective radicals most diluted. College general meetings are, in our experience, the most promising source of pressure to abolish grading; at such meetings radicals, armed with some of the overwhelming evidence that grading techniques are unreliable and vicious, can most easily and conspicuously defeat the academic conservatives—whose counter-arguments are usually anecdotal rather than statistical and are often very badly thought out. Departmental meetings, which are bound to happen, should follow and supplement general meetings.

For us, the aim of making universities happier and more sociable places is subordinate to that of bringing about such changes in society generally; we believe that the abolition of grading and the growth of co-operative work will help greatly as we move towards social revolution. We can expect support from people who do not share this wider purpose; they will think that they can use us, and we think we can use them. As for the academic conservatives, with their almost unsupported position their defeat over grading is inevitable, if radicals oppose them intelligently. The degree of conservative resistance will not determine whether grading ends—merely whether it ends with a bang or with a whimper.

NOTES TO SECTION 2

[1] Still, R. J., 'Psychological Illness Among Students In The Examination Period, Leeds University, mimeo, 1963.

[2] Proceedings of the British Student Health Association Conference of July 1968, p.161. Cf. Still, op. cit. and 'The Mental Health Of Students,' Leeds university, mimeo, 1966.

[3] E.g. Lucas, C. J., 'Examinations—The Forgotten Dimension', Times Ed. Supp., August 2, 1968.

[4] An article in the 'Universities Quarterly' of June 1967 tells of the "good-natured contempt" often expressed by academics towards student suffering over exams (Oppenheim, Jahoda, James, loc. cit., p.349).

[5] Conway, M., 'The Practitioner', June 1971, p.795.

[6] Malleson, N., 'A Handbook On British Student Health Services', 1965, p.62.

[7] Conway, art. cit.

[8] Still, 'The Mental Health Of Students,' p.10.

[9] Op. cit., pp.68f., and 'The Lancet' 1959 i p.225.

[10] The "Yerkes-Dodson Law"; Conway, art. cit. Cf. Still, op. cit., p.9, and Ryle, A., 'Student Casualties', 1969, p.49. Sprent, in Proceedings of the British Student Health Association Conference of 1964, p.93, refers to evidence that the skill of solving complex problems declines in each individual as anxiety increases.

[11] Daniels, M. J. M. and Schouten, J., 'The Screening Of Students', Council of Europe publication, 1970, pp.16ff. Cf. Ager. M., and Weltman, J., 'Universities Quarterly', June 1967, p.274.

[12] 'Examination Strain At Manchester University . . .', Manchester. 1966, mimeo. Cf. Ryle, loc. cit.

[13] 'Examination Strain At Manchester . . .' p.2. Cf. Still, 'The Mental Health Of Students', p.3. Concerning girls' performance in exams, it has been shown that those taking 'O' and 'A' Levels during or just before menstruation lost, on average, about 5% in their marks, and that special groups within those tested—e.g. those with a long menstrual cycle—lost on average even more; Dalton, K., 'The Lancet', 1968 ii, pp.1368ff.

[14] The pressure on overseas students is likely to have been increased by the massive increase in 1966 in the fees demanded from them.

[15] V. infra.

[16] Interestingly, this standard defence of exams is made more often orally than in the literature—doubtless because people usually think rather more carefully about what they commit to print. However, we have found one instance of this argument occurring in print: v. infra, 'A Conservative Voice', where it is quoted.

[17] Significantly, the more power that attaches to a job, the less likely the job is to have a special qualifying exam.

[18] Stengel, E., 'The Fear Of Examinations', 1960, p.13.

[19] Apparently it is sometimes felt that over-nervous candidates can by an act of will "snap out of" their condition, and consequently that for their nervous state they have "only themselves to blame". Medical writings, however, do not, so far as I know, provide any support for this feeling. Indeed, the general impracticality of the command to 'snap out of' conditions of anxiety and misery seems to be shown by the rarity of this verb's occurring in the past tense. We rarely or never hear it said that someone "snapped out of" such a condition—evidence that this process rarely or never takes place.

[20] See, e.g., Ryle, op. cit., p.101. Dr. C. J. Lucas tells me of a patient whom he once treated for severe exam panic, and yet who later reported no great anxiety at finding himself involved in the Greek ferry disaster off Brindisi in August 1971.

[21] The concentration in the following section on the statistics of male student suicide is not the result of male-chauvinism in the author, but is imposed by his material. Up to 1958 (the latest period for which anything like proper statistics has been published), the small number of women at British universities did not encourage statistical generalisations, and such statistics as exist show the suicide rate among women students to be much closer than that of men students to comparable non-student groups, and to have less apparent connection with exams; Carpenter, R. G., British Journal of Preventive and Social Medicine, 1959, p.173; Rook, A., British Medical Journal, 1959 i, pp.599ff.

[22] Carpenter, art. cit., pp.165-72; Ryle, op. cit., p.105. Contrast Cresswell, P. A. and Smith, G. A., 'Student Suicide' (pamphlet), 1968, p.8.

[23] Cresswell and Smith (op. cit.) argue for a significant correlation between student suicide rates and sex ratios—that (to state their thesis crudely) the universities with the most extremely disparate sex ratios tended to produce the highest suicide rates. They may well be right, tho' as they themselves observe—p.18—this would not preclude many suicides having had different or additional causes. The possibility of a causal connection between exams and suicide is not explored in their study.

[24] Art. cit., p.602.

[25] Carpenter, art. cit., p.170.

[26] E.g., M. Ross, 'Suicide Among College Students', American Journal of Psychiatry, August 1969, p.221, merely cites the 1959 Cambridge study by Rook, nor does Ryle in his book 'Student Casualties' (1969) cite any more recent study.

[27] Art. cit., p.170.

NOTES TO SECTION 3

[28] Universities Quarterly, 1970, p.143.

[29] Universities Quarterly, 1967, p.343.

[30] Daniels and Schouten, op. cit., p.12.

[31] See D. A. Allen, loc. cit.

[32] Op. cit., Appendix Two (A), Annex K.

[33] Cited in Daniels and Schouten, op. cit. pp.16f.

[34] Ibid.

[35] N. Entwhistle, 'New Academic', May 6, 1971.

[36] I. Kandel, 'Examinations and their Substitutes in the United States' cited in Cox, Universities Quarterly, 1967, p.304.

[37] Edgworth, cited in Cox, art. cit., p.295.

[38] Cox, ibid.

[39] W. C. Eells, Journal of Educational Psychology, 1930, pp.48-52.

[40] G. M. Bull, 'An Examination of the Final Examination in Medicine', in 'The Lancet' ii, 1956.

[41] Daniels and Schouten, op. cit., p.15.

[42] Bull, art. cit.

[43] Ibid.

[44] Stalnaker, cited in Cox, art. cit., p.303.

[45] D. M. Cast, British Journal of Educational Psychology, 1940.

[46] N.U.S. Executive Report on Examinations, 1969, p.14.

NOTES TO SECTION 4, SECTION 5, SECTION 6

47 J. Chadwick, 'Didaskalos', 1970, pp.274ff.

48 Ibid. p.274.

49 See Cox, art. cit., pp.306f.

50 Art. cit., p.275.

51 Ibid., p.274.

52 Ibid., p.275.

53 Ibid.

54 See, for example, Cox, Universities Quarterly, 1967, p.355.

55 v. supra, section 3.

56 Cox, Universities Quarterly, 1967, p.334.

57 P. Marris, 'The Experience of Higher Education', 1964; quoted in Cox, loc. cit.

58 Hubbard and Clemans, 'Multiple Choice Examinations in Medicine', 1961, cited in Cox, ibid.

59 Professor P. Edwards, quoted in 'The Observer', June 6, 1971.

60 R. C. Birney, Journal of Higher Education, February 1964, p.96.

61 An idea supported by an experiment at Essex university, which showed that the standard of work in term-time essays "went up dramatically" when they began to be used for final grading purposes; 'The Observer', June 6, 1971.

62 Dr. A. Ryle of the Sussex university Health Service, cited in 'The Observer', ibid.

Most of section 3 of this booklet was written by Brian Butterworth. The rest of the booklet was written by A. Powell. Like every study, our work depends heavily on the research and analysis done by earlier thinkers. We also owe much of what we have written to ideas, criticism and stimulation provided by our contemporaries at University College—notably, but not exclusively, by Mike Hennessy, Heather Sutton, Tim Cornell, Steve Ludlam, Professor Denys Holland and Dr. C. J. Lucas. The booklet is dedicated, in gratitude for much inspiration, to Steve Ludlam of Portsmouth and to the late Mary Ann Evans of Coventry.

Any orders for further copies of this booklet can usefully be addressed to:— A. Powell, Institute of Classical Studies, 31/4 Gordon Square, London WC1.

Printed by WILLIAM CROSS LTD., LEICESTER